CSC423: Data Analysis And Regression
CSC 324: Data Analysis & Statistical Software II

Week-9: Logistic Regression and Predictive Models for Qualitative Variables

---

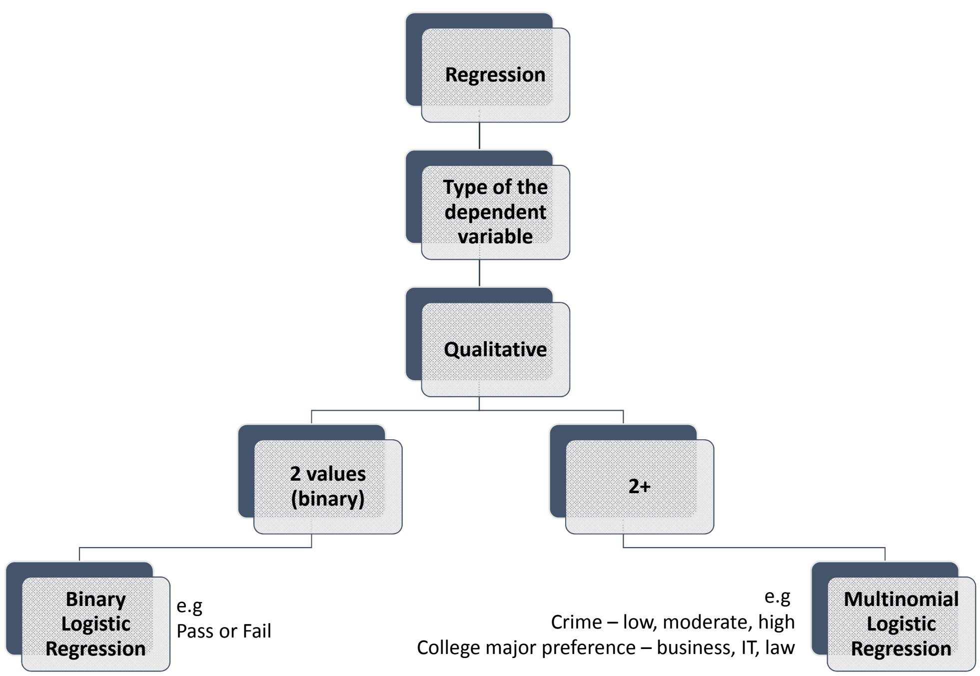
Outline

- Logistic regression models
  - One independent variable and Multiple independent variable
  - Model selection method and selection criteria
  - Assess overall goodness of fit
  - Diagnostics and residual analysis
  - Predictions

What is Logistic Regression?

- Logistic regression is a statistical method where one or more independent variables (IV) determine an outcome (DV)
- The outcome is measured with a dichotomous variable in which there are only two possible outcomes
- The Y (or DV) that is predicted in logistic regression is actually a probability, which ranges from 0 to 1 (binary)
- Example: logistic regression produces an equation that accurately predicts the probability of whether an individual will fall into either the
  - Pass or Fail category
  - Win or lose
  - Alive or dead
  - Healthy or sick
- Also known as logit regression, or logit model

Types of Logistic Regression



Applications of Logistic Regression

Customer reliability: bank wants to determine which customers are more likely to repay a loan, in relation to their income, credit score, loan amount, etc…

Response variable is binary:

Customer repaying loan: Y=Yes or No

Project risk analysis: probability that a project will be completed on time (or on budget) in relation to team experience, size, project requirements, etc…

Response variable is binary:

Project completed on time: Y=Yes or No

Students retention: probability of a student graduating in relation to sat scores, first year scores, attendance records, etc…

Response variable is binary:

Student graduating: Y=Yes or No

Example: Project Risk Analysis

How does programming experience affect the likelihood of completing a complex programming task within a specified timeframe?

25 programmers were given the same task. Their experience (in months) and the results of their success in completing the task are shown in the table

Person	Months of Experience	Task completion 1: Yes 0: No
1	14	0
2	29	0
3	6	0
…		
24	22	1
25	8	1

QUESTION:
Are programmers with more months of experience more likely to complete the task?

Example: Project Risk Analysis

- We can’t use a regression line to predict *Task Completion* based on *Months of Experience*, because the response variable is not quantitative (i.e. Yes/No)
- These problems are solved using Logistic Regression
- GOAL: Use data to analyze the relationship between months of experience and probability of completing the task

→ Odds of completing the task

Response variable ↓

Person	Months of Experience	Task completion 1: Yes 0: No
1	14	0
2	29	0
3	6	0
…		
24	22	1
25	8	1

What is Odds Ratio?

Probabilities range between 0 and 1

Let's say that the probability of completing a task is 0.8
→ p = 0.8

Then the probability of not completing a task is q = 1 - p
→ q = 1- (0.8) = 0.2

Odds is defined as the ratio of a probability that an event will occur divided by probability that event will not occur

$$\text{Odds (completing)} = \frac{p}{1-p} = \frac{0.8}{0.2} = 4$$

This means

- Odds of completing a task is 4 to 1
- Odds of not completing a task is 1 to 4 (i.e. 0.2/0.8 = 0.25 → 1/4)

Interpreting Odds Ratio

In other words, we can say that the odds that event Y = 1 occurs

$$\textbf{Odds} = \frac{p}{1-p} = \frac{P(Y=1)}{P(Y=0)}$$

Let p = Pr(Y=1) the probability of "completing" or of "success"

- If probability of completing a task is > 0.5

  If Pr(Y=1) **>** 0.5 → Odds > 1 e.g 0.6/(1-0.6) = 1.5

  => *higher chance of success*

- If probability of completing a task is = 0.5

  If Pr(Y=1) **=** 0.5 → Odds = 1 e.g 0.5/(1-0.5) = 1

  => *same chance of success or failure*

- If probability of completing a task is < 0.5

  If Pr(Y=1) **<** 0.5 → Odds < 1 e.g 0.4/(1-0.4) = 0.6

  => *higher chance of failure*

What is Odds Ratio?

Odds ratio ranges from 0 to ∞

Let's say that the probability of completing a task is 0
→ p = 0

Then the probability of not completing a task is q = 1 - p
→ 1-0 = 1

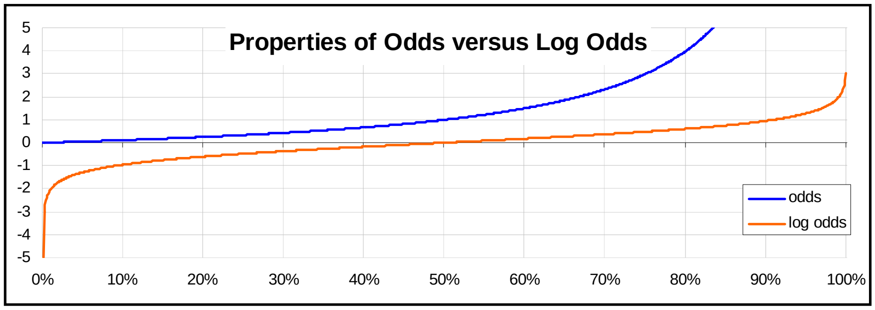
Odds = p/(1-p)
= 0/1
= 0

Let's say that the probability of completing a task is 1
→ p = 1

Then the probability of not completing a task is q = 1 - p
→ 1-1 = 0

Odds = p/(1-p)
= 1/0
= ∞

Graphical View: Odds vs Log Odds



Odds

- Not symmetric
- Skewed
- Varying from **0** to **∞**
- Is 1 when the probability is 50%

Log odds

- Is symmetric
- Approx. normal
- Varying from minus infinity to positive infinity, like a line
- Is 0 when the probability is 50%
- Is highly negative for low probabilities and highly positive for high probabilities

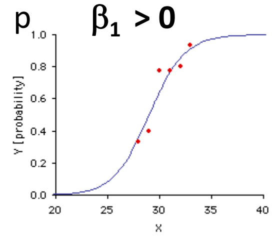
Logistic Regression Model

Simple case: Relationship between qualitative binary variable Y and one x-variable:

Model for probability p=Pr(Y=1) for each value x.

$$\log(\frac{p}{1-p}) = \beta_0 + \beta_1 x$$

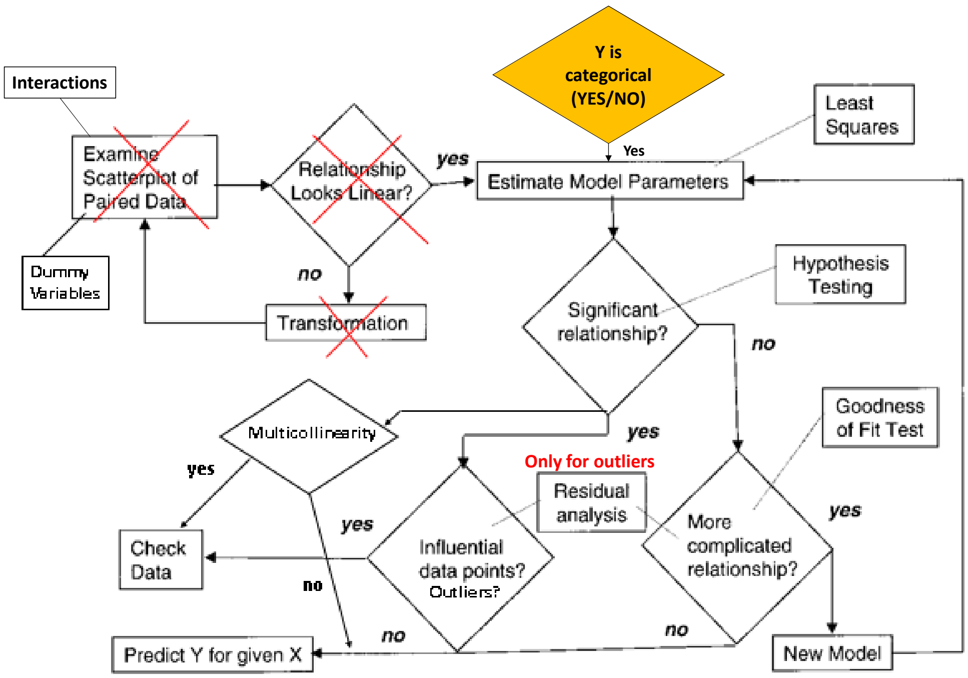
p β_1 **> 0**



The slope parameter measures the degree of association between the probability p =pr(Y=1) and the value of X

If β_1 **> 0**, (positive) then the odds of success **increases** with an increase in X
If β_1 **< 0**, (negative) then the odds of success **decreases** with an increase in X

How is a Logistic Regression Analysis done?



Example: Mississippi River Levee Failure

Factors relating to presence / absence of a levee failure at a site on middle Mississippi River. *(Levee - an embankment built to prevent the overflow of a river)*

<u>Variables/Columns</u>

- **Failure (Y) → 1=Yes, 0=No**
- Year
- River mile
- Sediments → 1=Yes, 0=No
- Borrow pit → 1=Yes, 0=No
- Meander location (winding course)→ 1=Inside bend, 2=outside bend, 3=chute, 4=straight
- Channel width
- Floodway width
- Constriction factor
- Land cover → 1=open water, 2=grassy, 3=agricultural, 4=forest
- Vegetation width
- Channel Sinuosity (curve/bending)
- Dredging intensity
- Bank Revetement (fencing)

k = 13
 4 – qualitative
 9 – quantitative
n = 70

Example: Mississippi River Levee Failure

1. Create the dataset in SAS

Obs	Failure	year	river_mile	sediments	borrow_pit	meander	channel_width	floodway_width	constriction_factor	land_cover	veg
1	1	1880	188.40	0	0	2	2512.91	6990.65	1.0000	1	
2	1	1908	190.00	1	0	1	1270.84	4343.60	3.0576	3	
3	1	1908	174.20	0	0	1	920.22	3395.81	1.0012	3	

…

* Logistic regression uses log odds, so you don't have to transform your Y variable

2. Create dummy variables

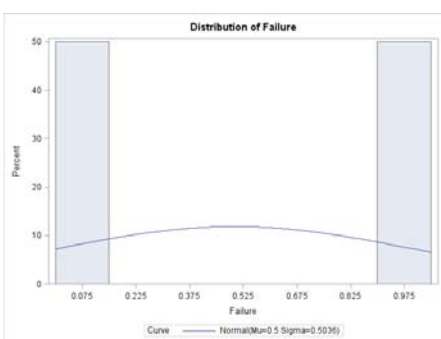
- Sediments → 1=Yes, 0=No
- Borrow pit → 1=Yes, 0=No

Already recoded, so no need for dummy variables

- Meander location → 1=Inside bend, 2=outside bend, 3=chute, 4=straight (base)
- Land cover → 1=open water, 2=grassy, 3=agricultural, 4=forest (base)

Example: Mississippi River Levee Failure

3. Histogram - Since "Failure" (Y) is binary – **not useful**

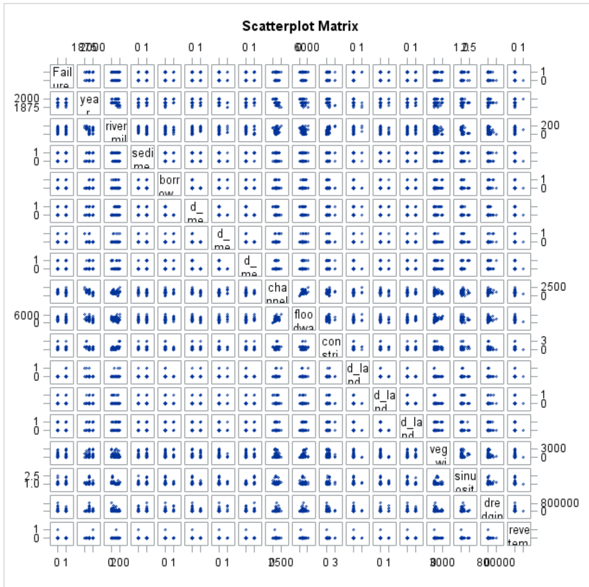


Quantiles (Definition 5)	
Level	Quantile
100% Max	1.0
99%	1.0
95%	1.0
90%	1.0
75% Q3	1.0
50% Median	0.5
25% Q1	0.0
10%	0.0
5%	0.0
1%	0.0
0% Min	0.0

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0	70	1	31
0	69	1	32
0	68	1	33
0	67	1	34
0	66	1	35

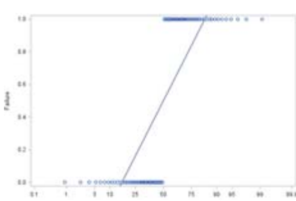
Example: Mississippi River Levee Failure

4. Scatterplot - Since “Failure” (Y) is binary – **not useful to see association between Y and x-var**



Association assumption doesn’t apply

Y is binary, therefore points are concentrated at 0 or 1



Example: Mississippi River Levee Failure

5. Pearson Correlation Coefficients - use to explore data and check for Multicollinearity not useful to check association between Y and x-var

Pearson Correlation Coefficients, N =
Prob > |r| under H0: Rho=0

	Failure	year	river_mile	sediments	borrow_pit	d_meander1	d_meander2	d_meander3	channel_width	floodway_width
Failure	1.00000	0.00000 1.0000	-0.02562 0.8333	0.37388 0.0014	-0.03925 0.7470	0.15587 0.1976	-0.10206 0.4005	0.12649 0.2967	-0.18089 0.1340	-0.03897 0.7487
year	0.00000 1.0000	1.00000	-0.14342 0.2362	0.12687 0.2953	0.31192 0.0086	0.02406 0.8433	-0.25461 0.0334	0.07759 0.5232	-0.57677 <.0001	-0.31494 0.0079
river_mile	-0.02562 0.8333	-0.14342 0.2362	1.00000	-0.33538 0.0045	-0.02191 0.8571	0.39872 0.0006	0.03790 0.7554	-0.51202 <.0001	-0.00907 0.9406	-0.13329 0.2713
sediments	0.37388 0.0014	0.12687 0.2953	-0.33538 0.0045	1.00000	-0.01016 0.9335	-0.16945 0.1608	-0.24069 0.0447	0.43654 0.0002	-0.10724 0.3769	-0.05912 0.6269
borrow_pit	-0.03925 0.7470	0.31192 0.0086	-0.02191 0.8571	-0.01016 0.9335	1.00000	-0.28267 0.0177	0.42867 0.0002	-0.01241 0.9188	-0.03627 0.7656	0.14732 0.2236
d_meander1	0.15587 0.1976	0.02406 0.8433	0.39872 0.0006	-0.16945 0.1608	-0.28267 0.0177	1.00000	-0.20045 0.0962	-0.41404 0.0004	-0.21136 0.0790	-0.17806 0.1403
d_meander2	-0.10206 0.4005	-0.25461 0.0334	0.03790 0.7554	-0.24069 0.0447	0.42867 0.0002	-0.20045 0.0962	1.00000	-0.19365 0.1082	0.37671 0.0013	0.21764 0.0703
d_meander3	0.12649 0.2967	0.07759 0.5232	-0.51202 <.0001	0.43654 0.0002	-0.01241 0.9188	-0.41404 0.0004	-0.19365 0.1082	1.00000	-0.01298 0.9150	0.15493 0.2003
channel_width	-0.18089 0.1340	-0.57677 <.0001	-0.00907 0.9406	-0.10724 0.3769	-0.03627 0.7656	-0.21136 0.0790	0.37671 0.0013	-0.01298 0.9150	1.00000	0.45964 <.0001
floodway_width	-0.03897 0.7487	-0.31494 0.0079	-0.13329 0.2713	-0.05912 0.6269	0.14732 0.2236	-0.17806 0.1403	0.21764 0.0703	0.15493 0.2003	0.45964 <.0001	1.00000
constriction_factor	0.00130 0.9915	-0.40176 0.0006	0.49828 <.0001	0.02292 0.8506	-0.20166 0.0941	0.09413 0.4383	-0.04584 0.7063	-0.08213 0.4991	0.19765 0.1010	0.03761 0.7572
d_land_cover1	0.21160 0.0787	-0.26375 0.0274	0.14127 0.2434	0.04666 0.7013	-0.09137 0.4519	0.16931 0.1612	0.18717 0.1208	-0.13383 0.2694	0.24797 0.0385	0.12995 0.2836
d_land_cover2	0.00000 1.0000	0.10834 0.3720	-0.00299 0.9804	-0.04132 0.7341	-0.03173 0.7943	0.25476 0.0333	0.05041 0.6786	-0.02840 0.8155	-0.21363 0.0758	-0.21861 0.0690
d_land_cover3	-0.15587 0.1976	0.20836 0.0835	0.23117 0.0542	-0.04393 0.7180	0.31693 0.0075	-0.08844 0.4666	-0.08909 0.4633	-0.06901 0.5703	-0.05315 0.6621	0.09688 0.4250
veg_width	-0.05631 0.6433	0.06666 0.5835	-0.18856 0.1180	0.02260 0.8527	0.15741 0.1931	-0.10622 0.3815	-0.13585 0.2622	0.03633 0.7653	-0.00016 0.9990	0.37152 0.0015
sinuosity	0.09588 0.4298	-0.02999 0.8053	-0.45285 <.0001	0.15024 0.2144	-0.06720 0.5805	-0.19197 0.1114	0.04467 0.7135	0.42271 0.0003	0.00213 0.9860	-0.05905 0.6273
dredging	-0.20223 0.0932	0.02993 0.8057	-0.31484 0.0079	0.00783 0.9487	0.11693 0.3350	-0.12780 0.2917	0.23365 0.0516	0.06298 0.6045	0.13616 0.2610	0.05780 0.6346
revetement	-0.12039 0.3209	0.17210 0.1543	0.03004 0.8050	-0.13503 0.2651	-0.05198 0.6691	-0.07881 0.5167	-0.03686 0.7619	-0.07614 0.5310	-0.10802 0.3734	-0.18067 0.1345

Pearson correlation → how change in one variable is related to change in another one

Y vs X-var

- Pearson correlation cannot deal with categorical variables (mostly because categorical variables don't have a notion of mean, which Pearson is based on)
- Binary (can be considered as continuous and calculate a *kind of* correlation
- This is clearly a hack, but it should work for simple exploration analysis

Example: Mississippi River Levee Failure

6. General Model Equation

$$\log\left(\frac{Failure=1}{Failure=0}\right) = \beta_0 + \beta_1\,\text{river\_mile} + \beta_2\,\text{sediments} + \beta_3\text{borrow\_pit} + \beta_4\text{d\_meander1} + \beta_5\text{d\_meander2} + \beta_6\text{d\_meander3} + \beta_7\text{channel\_width} + \beta_8\text{floodway\_width} + \beta_9\text{constriction\_factor} + \beta_{10}\text{d\_land\_cover1} + \beta_{11}\text{d\_land\_cover2} + \beta_{12}\text{d\_land\_cover3} + \beta_{13}\text{veg\_width} + \beta_{14}\text{sinuosity} + \beta_{15}\text{dredging} + \beta_{16}\text{revetement} + e$$

Example: Mississippi River Levee Failure

7. Fitting Full Model

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.8788	2.5297	0.1207	0.7283
river_mile	1	-0.00337	0.00967	0.1214	0.7275
sediments	1	1.8363	0.7551	5.9149	0.0150
borrow_pit	1	0.9661	1.2543	0.5933	0.4411
d_meander1	1	1.4352	0.8972	2.5588	0.1097
d_meander2	1	-0.9974	2.2840	0.1907	0.6624
d_meander3	1	0.1207	0.8913	0.0184	0.8922
channel_width	1	-0.00151	0.000959	2.4765	0.1156
floodway_width	1	0.000030	0.000312	0.0094	0.9227
constriction_factor	1	0.1713	0.7110	0.0581	0.8096
d_land_cover1	1	15.6634	333.5	0.0022	0.9625
d_land_cover2	1	-1.2262	1.1676	1.1029	0.2936
d_land_cover3	1	-1.0607	0.7757	1.8701	0.1715
veg_width	1	-0.00035	0.000464	0.5678	0.4511
sinuosity	1	0.1004	1.0967	0.0084	0.9271
dredging	1	-2.76E-6	2.195E-6	1.5851	0.2080
revetement	1	-12.8379	736.2	0.0003	0.9861

- Logistic Regression - Wald test is used to determine whether a certain predictor variable X is significant or not
- Exclude x-variables that are not significant (< 0.05)

$$\log\left(\frac{Failure=1}{Failure=0}\right) = 0.8788 + 1.8363\ \text{sediments}$$

Where sediments = 1 (present)
sediments = 0 (not present)

SAS: Example -Levee Failure – Model Statement

Logistic regression models are estimated in SAS using `PROC LOGISTIC` (similar syntax to `PROC REG`)

```
PROC LOGISTIC;
MODEL y-binary (event='1') = xvar1 xvar2…xvarn;
RUN;
```

Where y-binary is a variable that takes only values 0 and 1 – where 1 denotes "success"

SAS Documentation

PROC LOGISTIC Syntax
http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect003.htm

Example
http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect059.htm

SAS: Example -Levee Failure – Model Statement

```
* fit full logistic model;
proc logistic data=levee;
model Failure(event='1') = river_mile sediments borrow_pit
d_meander1 d_meander2 d_meander3 channel_width floodway_width
constriction_factor d_land_cover1 d_land_cover2 d_land_cover3
veg_width sinuosity dredging revetement;
run;
```

event='1' → the levee will fail / breach

Estimation Procedure for Logistic Regression

- Parameter estimates are computed using Maximum Likelihood Estimation (MLE)
- The inference for logistic regression models is similar to standard linear regression
- They become unbiased minimum variance estimators as the sample size increases
- They have approximate normal distributions and approximate sample variances that can be calculated and used to generate confidence bounds
- Likelihood functions can be used to test hypotheses about models and parameters

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.8788	2.5297	0.1207	0.7283
sediments	1	1.8363	0.7551	5.9149	0.0150

Maximum Likelihood Estimation

- Computes the parameter values that maximize the probability function of Y given the data (called **likelihood function**):

$$\max_{\beta_0, \beta_1, \ldots, \beta_k} \Pr(Y \mid \beta_0, \beta_1, \ldots, \beta_k, data)$$

- MLE's of logistic regression model are found using numerical optimization algorithms
- MLE's properties allow us to compute significance tests on model parameters and diagnostics

Tests for a Single Parameter β_i

The significance test on model parameters β that evaluates the influence of x-variables on p is:

H_0 : β_i=0 *(x-variable has no effect on pr(Y))*

H_a : β_i<>0 *(x-variable influences pr(Y))*

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.8788	2.5297	0.1207	0.7283
river_mile	1	-0.00337	0.00967	0.1214	0.7275
sediments	1	1.8363	0.7551	5.9149	0.0150
borrow_pit	1	0.9661	1.2543	0.5933	0.4411

- It is computed using the **Wald test statistic** $$z = \frac{\hat{\beta}_i}{s.e.(\hat{\beta}_i)}$$
- Z has approximately a normal distribution N(0,1)
- **SAS reports** equivalent test that uses the **chi-square statistic z^2**
- **When the null hypothesis is true**, z^2 has a distribution that is approximately a *chi-square distribution with 1 degree* of freedom

Interpret test p-values as usual: Small p-values (<0.05) provide strong evidence that the null hypothesis can be rejected, and therefore corresponding x-variable should be kept in the model

Example: Levee Failure - Interpreting Parameter Estimates

The predictive model for *p = Pr(Y=Failure)* is estimated as:

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

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.8788	2.5297	0.1207	0.7283
sediments	1	1.8363	0.7551	5.9149	0.0150

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 0.8788 + 1.8363 \text{ sediments}$$

Where sediments=1 (present)
sediments = 0 (not present)

What does the slope β_1 = 1.8363 mean?

Log odds log(p/(1-p)) of Failure increase by 1.8363, for when sediments are present (i.e. =1)

Using the anti-log function exp(1.8363) = 6.27. The odds p/(1-p) of Failure increases by 527%, when sediments = 1 → i.e. [(6.27-1)*100]

Note: e^{β_1} -1 is the percentage change in odds of success for every unit increase in X, holding all the other x fixed

Model Selection Methods and Selection Criteria

Model Selection Methods –

Backward, Forward or Stepwise selection procedure: (SAS)

- Similar to regression analysis
- Removes (backward) or adds (forward, stepwise) one variable at the time, by eliminating the variable with the large p-value based on the Wald statistic

AIC and BIC procedure: (for other software)

- Similar to regression analysis

Model Selection Criteria –

- AIC - smallest AIC is most desirable
- SC - smallest SC is most desirable

AIC and **SC** penalizes for the number of insignificant predictors in the model

AIC/SC: http://www.ats.ucla.edu/stat/sas/output/sas_logit_output.htm

Example: Mississippi River Levee Failure

8. Model Selection – Same methods as linear regression

```
* fit logistic model, and run stepwise selection procedure;
proc logistic data=levee;
model Failure(event='1') =  river_mile sediments borrow_pit
d_meander1 d_meander2 d_meander3 channel_width floodway_width
constriction_factor d_land_cover1 d_land_cover2 d_land_cover3
veg_width sinuosity dredging revetement
     /selection = stepwise;
run;
```

Stepwise, forward, or backward methods only

Example: Mississippi River Levee Failure

8. Model Selection – Same methods as linear regression

Stepwise, and Backward methods

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.8788	2.5297	0.1207	0.7283
sediments	1	1.8363	0.7551	5.9149	0.0150

Sediments = 1 (Yes)

Forward method

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.4487	0.5214	7.7204	0.0055
sediments	1	1.9082	0.5858	10.6126	0.0011
d_meander1	1	1.2435	0.6349	3.8362	0.0502

Sediments = 1 (Yes)
d_meander1 → 1=Inside bend

Goodness-of-Fit Test: Likelihood Ratio (LR) Test

- LR test is similar to F-test in linear regression. It is used to compare two models: a model M1 (with predictors) and a simpler model M0 (no predictors)

- Over all goodness of fit can used to test whether certain model parameters are zero by comparing the log likelihood for the fitted model M1 with the log likelihood for a simpler model M0

- For example you want to compare the hypotheses:

  H_0: $\beta_1=\beta_2=0$ (*hypothesis of all parameters=0 corresponds to an "empty" model or M0: logit(p)=* β_0)

  H_a: all $\beta_i \neq 0$ (*hypothesis corresponds to model with some covariates or Model M1: logit(p)* = $\beta_0 + \beta_1 x_1 + \beta_2 x_2$)

Logit(p) - alternative terms log odds or *log(p/(1-p)*

Covariates – alternative terms explanatory variable, independent variable, or predictor

Goodness-of-Fit Test: Likelihood Ratio (LR) Test

- Hypotheses:

  H_0: $\beta_1=\beta_2=0$ (*hypothesis of all parameters=0 corresponds to an "empty" model or M0: logit(p)= β_0*)

  H_a: all $\beta_i \neq 0$ (*hypothesis corresponds to model with some covariates or Model M1: logit(p) = $\beta_0 + \beta_1 x_1 + \beta_2 x_2$*)

- The test statistic is LR statistic =
  $-2\,(L_0 - L_1) = -2\,[\log(\Pr(Y|M0)) - \log(\Pr(Y|M1))]$

- The test statistic has approximately χ^2 distribution (computed by statistical software). When LR is large, then M1 is a better choice and provides a better fit than M0

LR Statistic: https://en.wikipedia.org/wiki/Likelihood-ratio_test

Chi-squared Distribution: http://en.wikipedia.org/wiki/Chi-squared_distribution

Example: Levee Failure - Goodness-of-Fit Test i.e. Likelihood Ratio (LR) Test

- Hypotheses

  H_0: $\beta_1=\beta_2=0$ (*hypothesis of all parameters=0 corresponds to an "empty" model or M0: logit(p)= β_0*)

  H_a: all $\beta_i \neq 0$ (*hypothesis corresponds to model with some covariates or Model M1: logit(p) = $\beta_0 + \beta_1 x_1 + \beta_2 x_2$*)

Stepwise, and backward methods

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	10.0412	1	0.0015
Score	9.7849	1	0.0018
Wald	9.2607	1	0.0023

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.8788	2.5297	0.1207	0.7283
sediments	1	1.8363	0.7551	5.9149	0.0150

LR is high and the p-value associated with the LR is almost zero.

Reject H_0 .

This model is better than the null model

Example: Levee Failure – Comparing 2 Models

Selection Method

Stepwise, and backward methods

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	99.041	90.999
SC	101.289	95.496
-2 Log L	97.041	86.999

Selection Criteria

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	10.0412	1	0.0015
Score	9.7849	1	0.0018
Wald	9.2607	1	0.0023

GOF

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.8788	2.5297	0.1207	0.7283
sediments	1	1.8363	0.7551	5.9149	0.0150

Std. Err. Of x-var Sig. x-var

Good Model minimizes AIC and SC

Used just for comparing, but to select the best model, need to check additional things

Forward methods

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	99.041	88.777
SC	101.289	95.523
-2 Log L	97.041	82.777

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	14.2634	2	0.0008
Score	13.2485	2	0.0013
Wald	11.1592	2	0.0038

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.4487	0.5214	7.7204	0.0055
sediments	1	1.9082	0.5858	10.6126	0.0011
d_meander1	1	1.2435	0.6349	3.8362	0.0502

How to read Logistic Regression SAS o/p: http://www.ats.ucla.edu/stat/sas/output/sas_logit_output.htm

Example: Levee Failure – Most Influential Predictor

Stepwise, and backward methods

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-0.8938	0.3957	5.1027	0.0239	
sediments	1	1.5870	0.5215	9.2607	0.0023	0.4377

Forward methods

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate
Intercept	1	-1.4487	0.5214	7.7204	0.0055	
sediments	1	1.9082	0.5858	10.6126	0.0011	0.5263
d_meander1	1	1.2435	0.6349	3.8362	0.0502	0.3164

Diagnostics…

1. Binary logistic regression requires the DV to be binary (1,0)
2. Since logistic regression assumes that P(Y=1) is the probability of the event occurring, it is necessary that the DV is coded accordingly. i.e., for a binary regression, the factor level 1 of the dependent variable should represent the desired outcome
3. It requires quite large sample sizes. Because maximum likelihood estimates are less powerful than ordinary least squares
   - At least 10 cases per independent variable, some statisticians recommend at least 30 cases for each parameter to be estimated
   - Make sure it has enough observations for each case (1 & 0)

Note:

If there isn't enough sample or there are many cells with no response, parameter estimates and standard errors are likely to be unstable and maximum likelihood estimation (MLE) of parameters could be impossible to obtain

Diagnostics…

4. Model should have little or no multicollinearity. If multicollinearity is present centering the variables might resolve the issue. If this does not lower the multicollinearity, a factor analysis with orthogonally rotated factors should be done before the logistic regression is estimated

5. Model should have no outliers or significant influential points.
   - Outliers → Use Pearson or Deviance residual close to or exceeding ± 3
   - Influential Points → Use Dfbetas

Residuals

Residual analysis is more difficult than the linear regression models

Pearson residuals $r_{pi} = \dfrac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$

- Difference between observed and fitted values and divide by an estimate of the standard deviation of the observed value
- Observations with a Pearson residual close to or exceeding ± 3 may be worth a closer look → Outliers

Deviance residuals *(more complicated standardization),* but with properties similar to least squares residuals

- An alternative residual, based on the deviance or likelihood ratio chi-squared statistic
- Observations with a deviance residual close to or exceeding ± 3 may indicate lack of fit → outliers

Studentized Pearson residuals *(more complicated standardization)*

Regression Diagnostics for Binary Data: http://data.princeton.edu/wws509/notes/c3s8.html

Example: Levee Failure – Diagnostics…

1. Binary logistic regression requires the DV to be binary (1,0)
2. Since logistic regression assumes that P(Y=1) is the probability of the event occurring, it is necessary that the DV is coded accordingly. i.e., for a binary regression, the factor level 1 of the dependent variable should represent the desired outcome

Obs	Failure
1	1
2	1
3	1
4	1
67	0
68	0
69	0
70	0

3. It requires quite large sample sizes. Because maximum likelihood estimates are less powerful than ordinary least squares

   k = 16 (Predictors - w/dummy)

   n = 70 (y = 1 → 35 obs ; y = 0 → 35 obs)

   Not large enough sample Should have at least 160 obs

Example: Levee Failure - Diagnostics…

4. Model should have little or no multicollinearity

- Multicollinearity occurs when x-variables are strongly correlated with each other
- Similar to regression analysis, it causes computational problems and inflates standard error of estimates

Estimated Correlation Matrix			
Parameter	Intercept	sediments	d_meander1
Intercept	1.0000	-0.7961	-0.6132
sediments	-0.7961	1.0000	0.3925
d_meander1	-0.6132	0.3925	1.0000

No sign of multicollinearity

You can run this on the full model to detect collinearity before model selection

```
*Final Model - based on forward method;
proc logistic data=levee;
model Failure(event='1') = sediments d_meander1/corrb;
run;
```

Generates Correlation matrix for betas

SAS: PROC CORR vs CORRB option in Logistic Regression

PROC CORR (Pearson correlation coefficient) gives you the correlation of the variables, while CORRB (option provided with logistic regression model statement) is the correlation of the coefficients of these variables in the model

Pearson Correlation Coefficients, N = 70 Prob > |r| under H0: Rho=0			
	Failure	sediments	d_meander1
Failure	1.00000	0.37388 0.0014	0.15587 0.1976
sediments	0.37388 0.0014	1.00000	-0.16945 0.1608
d_meander1	0.15587 0.1976	-0.16945 0.1608	1.00000

Doesn't change with addition or removal of predictors

Not the same → Based on parameter estimates. Each time a predictor is added, estimates change, hence the correlation matrix value

Estimated Correlation Matrix			
Parameter	Intercept	sediments	d_meander1
Intercept	1.0000	-0.7961	-0.6132
sediments	-0.7961	1.0000	0.3925
d_meander1	-0.6132	0.3925	1.0000

Estimated Correlation Matrix					
Parameter	Intercept	sediments	d_meander1	d_meander2	d_meander3
Intercept	1.0000	-0.6512	-0.7090	-0.5155	-0.3416
sediments	-0.6512	1.0000	0.3139	0.2734	-0.2749
d_meander1	-0.7090	0.3139	1.0000	0.3815	0.3699
d_meander2	-0.5155	0.2734	0.3815	1.0000	0.2299
d_meander3	-0.3416	-0.2749	0.3699	0.2299	1.0000

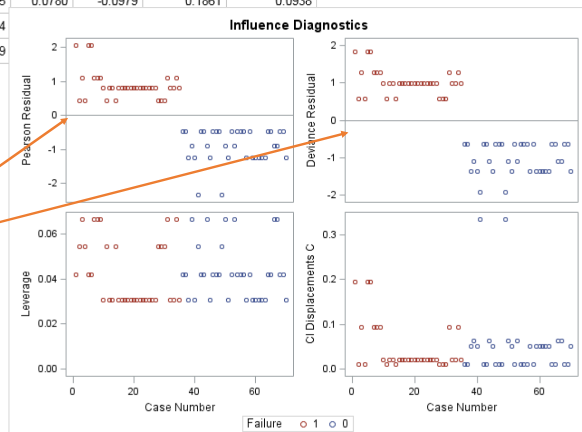
See post:
http://stackoverflow.com/questions/6172589/how-to-read-the-correlation-matrix-output-by-proc-logistic-and-proc-reg-in-sas

Example: Levee Failure - Diagnostics…

5. Model should have no outliers. Use Pearson or Deviance residual close to or exceeding ± 3

Regression Diagnostics										
	Covariates									Confidence Interval Displacement
Case Number	sediments	d_meander1	Pearson Residual	Deviance Residual	Hat Matrix Diagonal	Intercept DfBeta	sediments DfBeta	d_meander1 DfBeta	Confidence Interval Displacement C	
1	0	0	2.0634	1.8219	0.0419	0.4407	-0.3508	-0.2702	0.1942	
2	1.0000	1.0000	0.4268	0.5785	0.0544	-0.0545	0.0684	0.0888	0.0111	
3	0	1.0000	1.1081	1.2657	0.0665	0.0780	-0.0979	0.1861	0.0938	
4	1.0000	1.0000	0.4268	0.5785	0.0544					
5	0	0	2.0634	1.8219	0.0419					

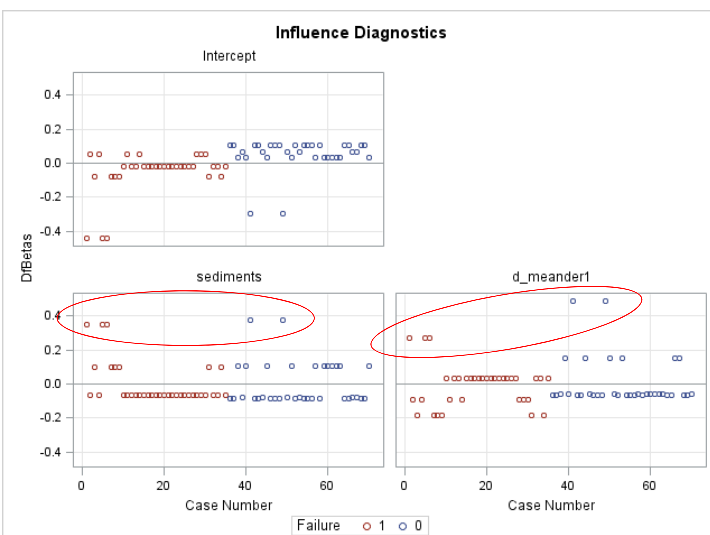
Plots and the table don't show any observations that have a Pearson or Deviance residual close to or exceeding ± 3



Example: Levee Failure - Diagnostics…

5. Model should have no significant influential points

Dfbetas – **USE: Check case when |Dfbeta| > 2/√n**



n= 70

|Dfbeta| > 2/sqrt(n)
|Dfbeta| > 0.23

Obs. 1, 5, 6, 41 & 49

SAS: Example: Levee Failure – Diagnostics…

```
ods graphics on;
* Check...;
* std. coefficients and r-squared;
* Final Model - based on forward method;
proc logistic data=levee;
model Failure(event='1') = sediments d_meander1
       /influence iplots corrb stb ;
run;
ods graphics off;
```

Standardized Coefficients

Influential Obs.

Check for Outliers (Residual plots)

Correlation matrix (Collinearity)

Logistic Regression – Model Statement Options:
https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect016.htm

Logistic Regression – Example:
https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_logistic_sect057.htm

Computing Predicted Probabilities

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

To obtain prediction estimates for values x_1 and x_2, the logit equation is solved for *p*

$$\hat{p} \mid x_1, x_2 = \frac{1}{1 + e^{-(b0 + b1x1 + b2x2)}}$$

Example: Levee Failure - Predicted Probabilities

Predicted probability value and confidence interval of <u>having sediments</u> when the <u>meander location is on the 'inside bend'</u>

Obs	sediments	d_meander1	d_land_cover3	_FROM_	_INTO_	IP_0	IP_1	_LEVEL_	phat	lcl	ucl
1	1	1	.	.	1	0.15408	0.84592	1	0.84592	0.60736	0.95119

Predicted probability = 0.845 ← no need to transform

95% confidential interval is (0.607, 0.951) → 2 ways to specify

(1) 95% of the time, the predicted probability will fall within 0.607 and 0.951

OR

(2) The corresponding 95% confidence limits for the odds ratio are

[exp(0.607) – 1]* 100, [exp(0.951) – 1] * 100

The odds of having the levee fail when sediments are present and when the meander location is in the inside bend will increase between 83.4% and 158.8%

The LOGISTIC Procedure: http://www.math.wpi.edu/saspdf/stat/chap39.pdf

SAS: Example – Levee Failure

```
data new;
input sediments d_meander1;
datalines;
1 1
;
data pred;
set new levee;
run;

* logistic regression model;
proc logistic data=pred;
model Failure(event='1')=sediments d_meander1;
output out=pred p=phat lower=lcl upper=ucl
             predprobs=(individual);
run;
```

phat → predicted probabilities (phat), lower and upper prediction

lcl and ucl →intervals lower and upper CI

predprob=(individual) → produces predicted probabilities for each response level

The LOGISTIC Procedure: http://www.math.wpi.edu/saspdf/stat/chap39.pdf