巨量資料與統計分析

政治大學統計系余清祥

2024年10月22日

第六週：廣義線性模式

http://csyue.nccu.edu.tw

數量分析

■透過數理模型描述觀察結果：

觀察現象 = 模型 + 誤差

或是

$y = f(x) + \text{error}$；觀察值 = 訊號 + 雜訊。

■數量化模型的關鍵：

→量化目標值 y：定義問題！

→選取關鍵變數：$x_1, x_2, \ldots, x_p$

→建立量化模型：統計學習、機器學習。

廣義線性模型(Generalized Linear Model)

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

Simple Linear Regression

Model the mean of a numeric response Y as a function of a single predictor X, i.e.

$$E(Y|X) = b_o + b_1 f(x)$$

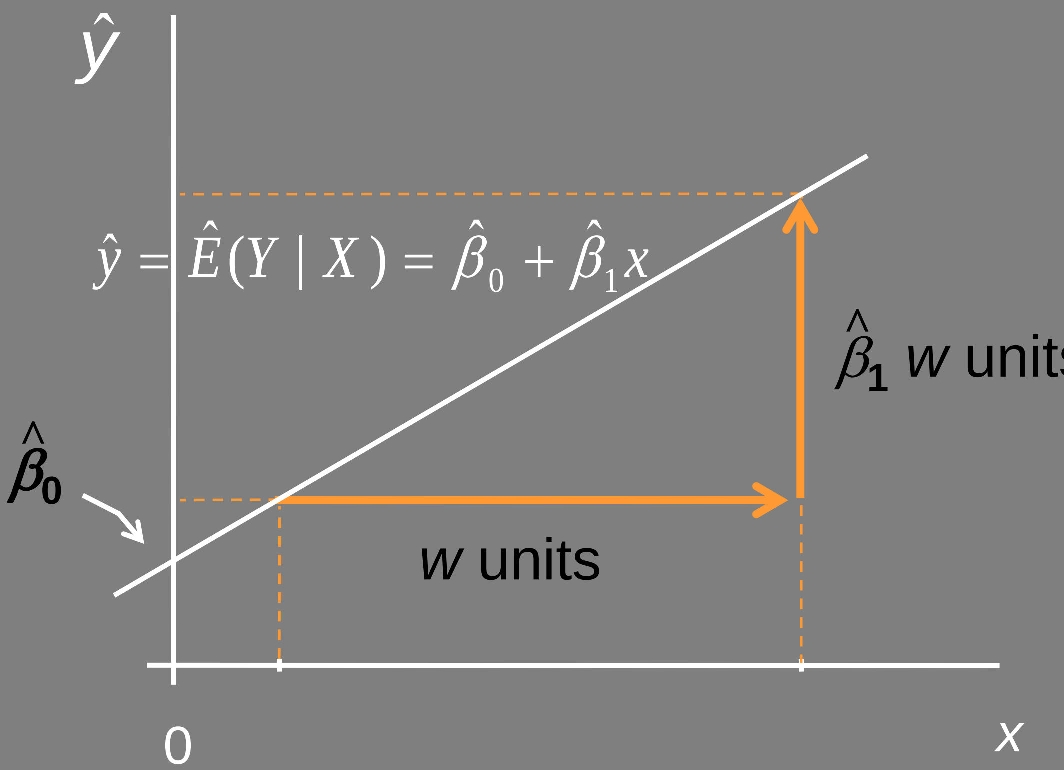
Here $f(x)$ is any function of X, e.g.

$f(x) = X$ ➔ $E(Y|X) = b_o + b_1 X$ (line)

$f(x) = ln(X)$ ➔ $E(Y|X) = b_o + b_1 ln(X)$ (curved)

The key is that E(Y|X) is a linear in the parameters b_o and b_1 but not necessarily in X.

Simple Linear Regression



$\hat{\beta}_0$ **= Estimated Intercept**

= $\hat{y}$-value at $x = 0$

Interpretable only if x = 0 is a value of particular interest.

$\hat{\beta}_1$ **= Estimated Slope**

= Change in $\hat{y}$ for every unit increase in x

= estimated change in the mean of Y for a unit change in X.

Always interpretable!

Multiple Linear Regression

We model the mean of a numeric response as linear combination of the predictors themselves or some functions based on the predictors, i.e.

$$E(Y|\mathbf{X}) = \beta_o + \underbrace{\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p}$$

Here the terms in the model are the predictors

$$E(Y|\mathbf{X}) = \beta_o + \underbrace{\beta_1 f_1(\mathbf{X}) + \beta_2 f_2(\mathbf{X}) + \ldots + \beta_k f_k(\mathbf{X})}$$

Here the terms in the model are ***k*** different functions of the ***p*** predictors

Multiple Linear Regression

For the classic multiple regression model

$$E(Y|\mathbf{X}) = b_o + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

the regression coefficients (b_i) represent the estimated change in the mean of the response Y associated with a unit change in X_i while the other predictors are held constant.

They measure the association between Y and X_i **adjusted** for the other predictors in the model.

General Linear Models

- Family of regression models
- Response / Model Type

Response	Model Type
– Continuous	Linear regression
– Counts	Poisson regression
– Survival times	Cox model
– Binomial	Logistic regression

- Uses
  - Control for potentially confounding factors
  - Model building , risk prediction

Logistic regression

- Most important model for ***categorical response*** (y_i) data
- Categorical response with 2 levels (*binary*: 0 and 1)
- Categorical response with ≥ 3 levels (nominal or ordinal)
- Predictor variables (x_i) can take on *any* form: binary, categorical, and/or continuous

Logistic Regression

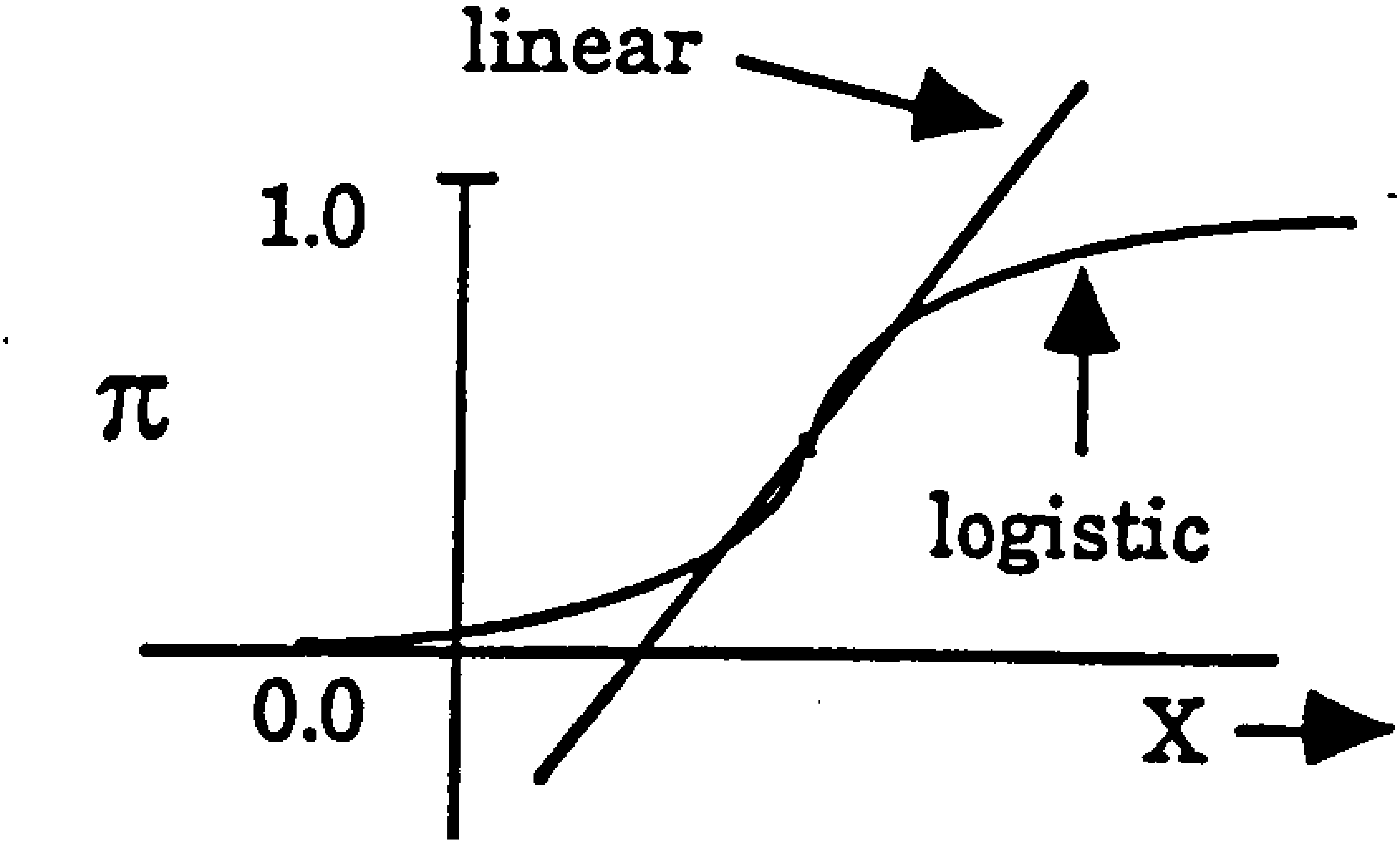
- Models relationship between set of variables X_i
  - dichotomous (yes/no, smoker/nonsmoker,…)
  - categorical (social class, race, ...)
  - continuous (age, weight, gestational age, ...)

and

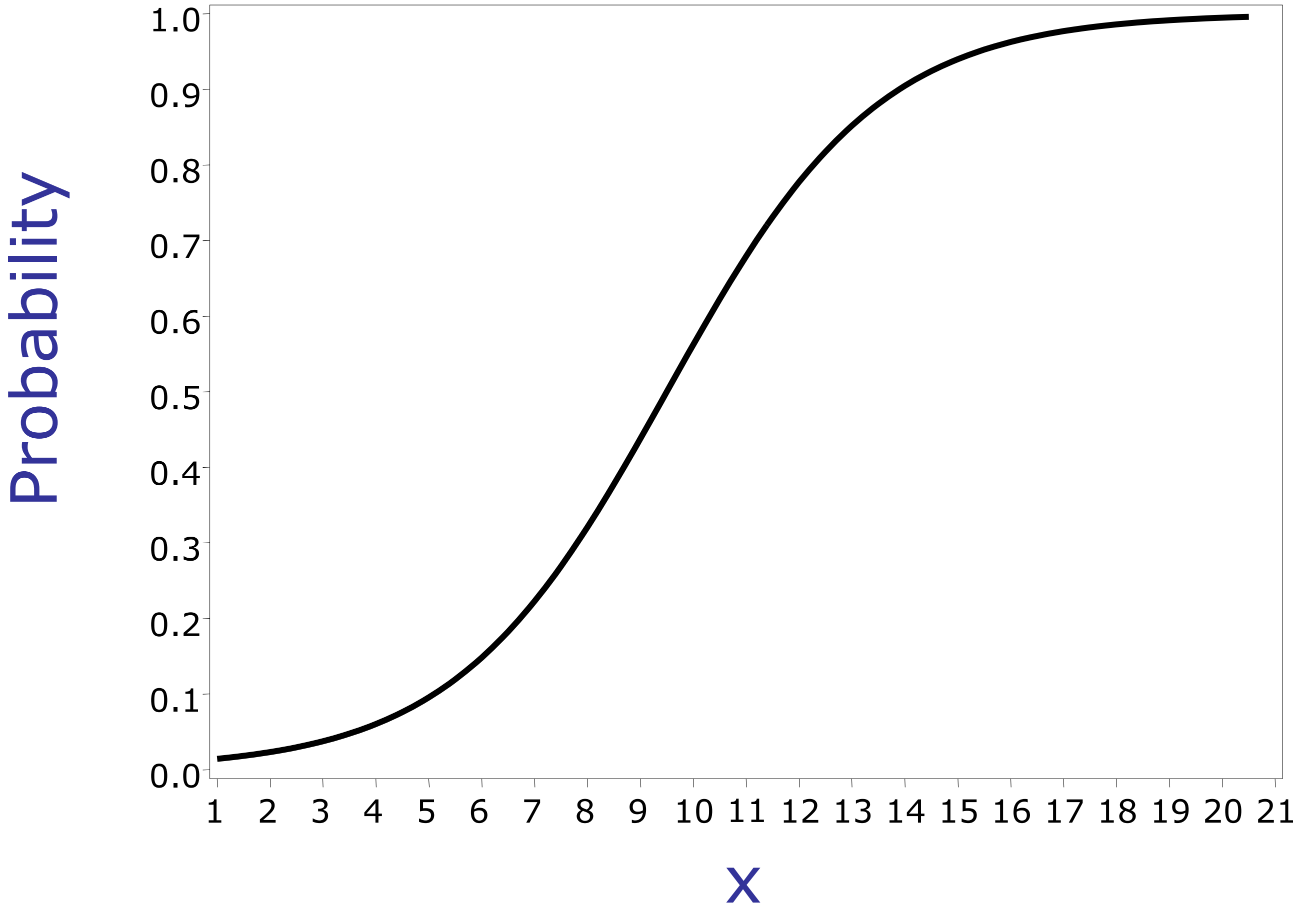
  - dichotomous categorical response variable Y

  e.g. Success/Failure, Remission/No Remission, Survived/Died, CHD/No CHD, Low Birth Weight/Normal Birth Weight…

Sigmoid curve for logistic regression



Logistic Regression Curve



Logit Transformation

Logistic regression models transform probabilities called *logits*.

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

where

i indexes all cases (observations).

p_i is the probability the event (a sale, for example) occurs in the i^{th} case.

log is the natural log (to the base e).

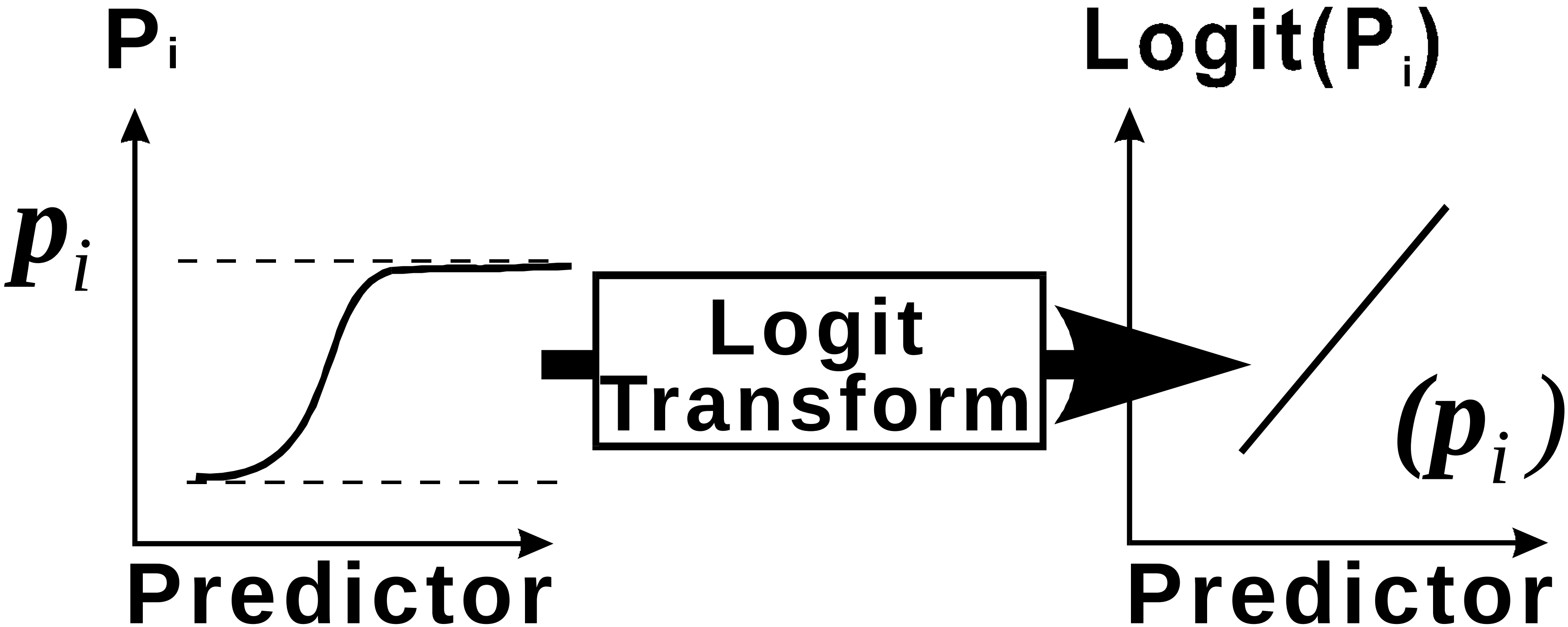
Logistic regression model with a single continuous predictor

$$\text{logit}\ (p_i) = \log\ (\text{odds}) = \beta_0 + \beta_1 X_1$$

where

logit(p_i)	logit transformation of the probability of the event
β_0	intercept of the regression line
β_1	slope of the regression line

Assumption



Interpretation of a single *continuous* parameter

- The sign (±) of β determines whether the **log odds** of y is increasing or decreasing *for every 1-unit increase* in x.
- If $\beta > 0$, there is an increase in the **log odds** of y for every 1-unit increase in x.
- If $\beta < 0$, there is a decrease in the **log odds** of y for every 1-unit increase in x.
- If $\beta = 0$ there is *no linear relationship* between the **log odds** and x.

Parameter interpretation (ctd).

- Exponentiating both sides of the logit link function we get the following:

$$\left(\frac{p_i}{1-p_i}\right) = \text{odds} = \exp(\beta_0 + \beta_1 X_1) = e^{\beta 0}\, e^{\beta 1 X1}$$

- The odds increase **multiplicatively** by e^{β} for every 1-unit increase in x.
- Whether the increase is greater than 1 or less than one depends on whether $\beta > 0$ or $\beta < 0$.
- The odds at $X = x+1$ are e^{β} times the odds at $X = x$. Therefore, **e^{β} is an odds ratio!**

Logistic regression model with a single *categorical ($\geq$ 2 levels)* predictor

$$logit\ (p_i) = log\ (odds) = \beta_0 + \beta_k X_k$$

where

$logit(p_i)$	logit transformation of the probability of the event
β_0	intercept of the regression line
β_k	difference between the logits for category k vs. the reference category

Logistic Regression

Example: Coronary Heart Disease (CD) and Age

In this study sampled individuals were examined for signs of CD (present = 1/absent = 0) and its potential relationship with the age (yrs.) was considered.

	Agegrp	Age	CD	Agegrp	Age	CD
1	1	20	0	2	30	0
2	1	23	0	2	30	0
3	1	24	0	2	30	0
4	1	25	0	2	30	0
5	1	25	1	2	30	1
6	1	26	0	2	32	0
7	1	26	0	2	32	0
8	1	28	0	2	33	0
9	1	28	0	2	33	0
10	1	29	0	2	34	0
11	2	30	0	2	34	0

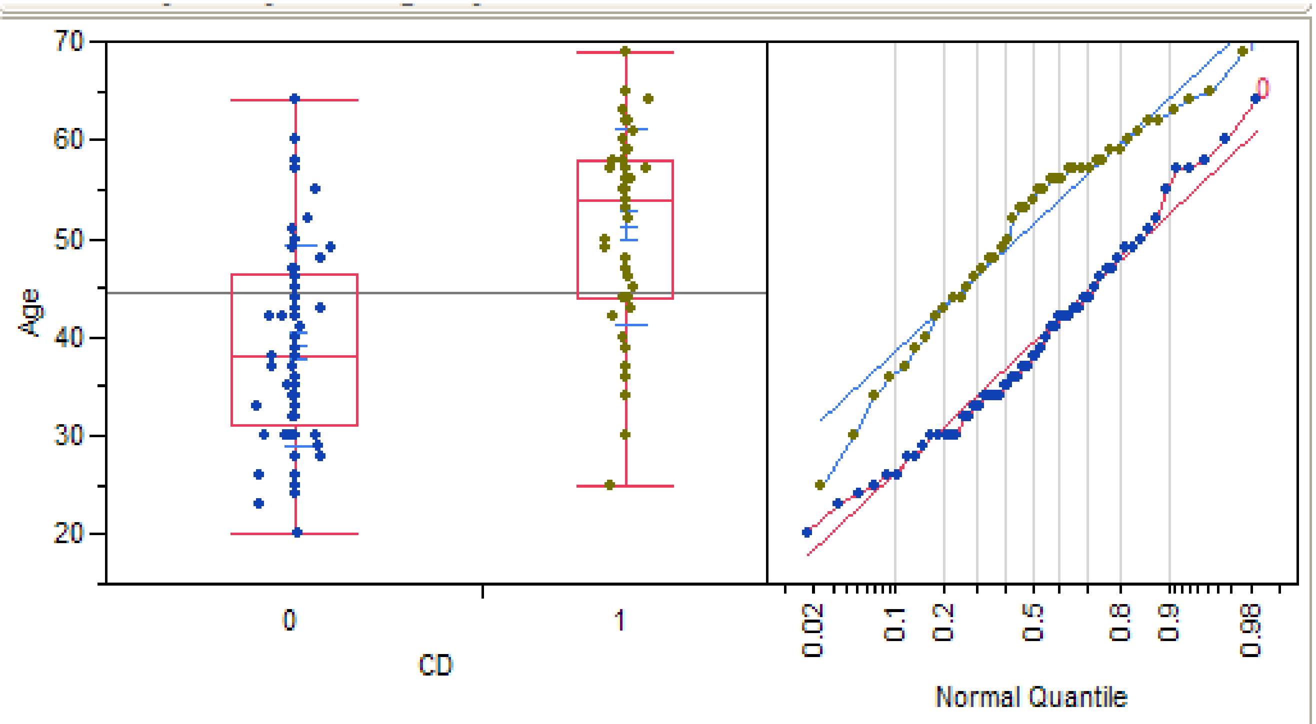
• • •

Agegrp	Age	CD
8	60	0
8	60	1
8	61	1
8	62	1
8	62	1
8	63	1
8	64	0
8	64	1
8	65	1
8	69	1

Note: This is a portion of the raw data for the 100 subjects who participated in the study.

Logistic Regression

- How can we analyze these data?



Means and Std Deviations

Level	Number	Mean	Std Dev	Std
0	57	39.1754	10.2018	
1	43	51.2791	9.9793	

t Test

1-0

Assuming unequal variances

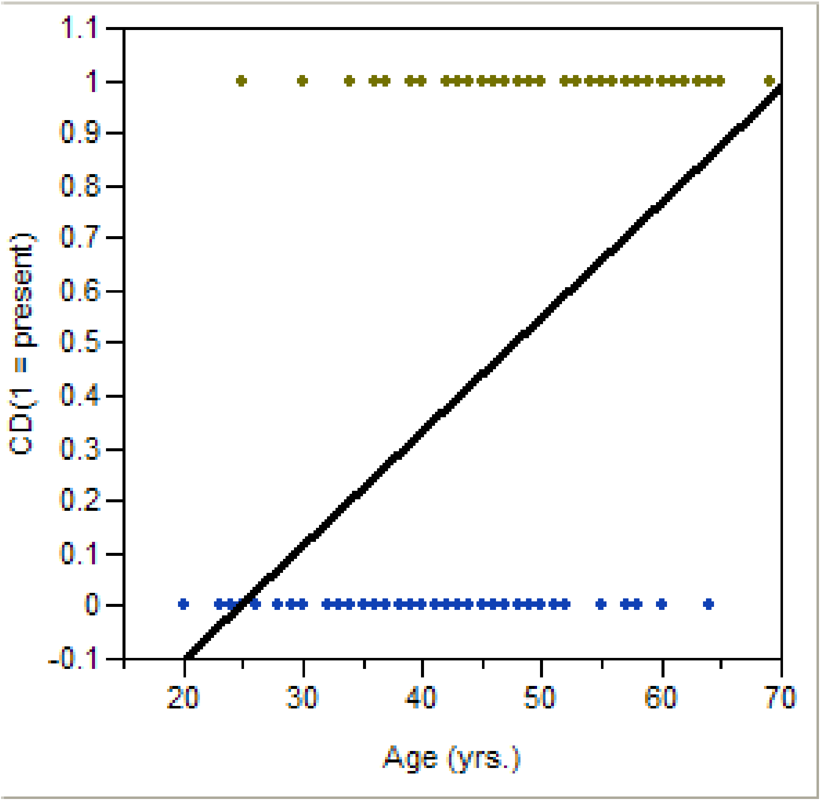
Difference	12.1036	t Ratio	5.94727
Std Err Dif	2.0352	DF	91.61987
Upper CL Dif	16.1459	Prob > |t|	<.0001*
Lower CL Dif	8.0614	Prob > t	<.0001*
Confidence	0.95	Prob < t	1.0000

Non-pooled t-test

The mean age of the individuals with some signs of coronary heart disease is 51.28 years vs. 39.18 years for individuals without signs ($t = 5.95$, $p < .0001$).

Logistic Regression

Simple Linear Regression?

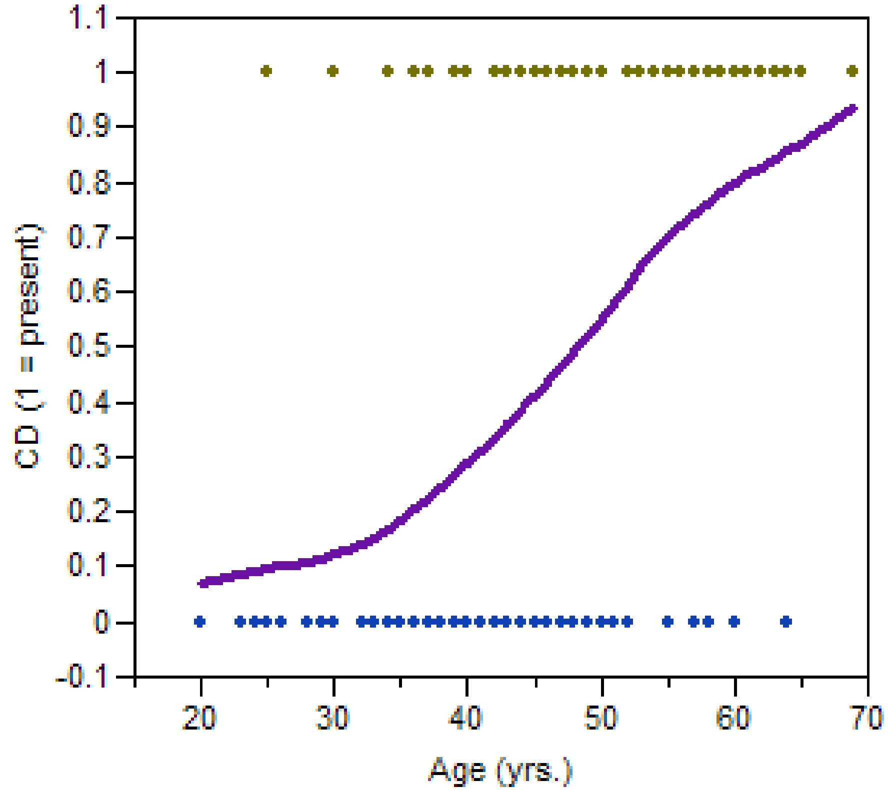


$E(CD \mid Age) = -.54 + .02 \cdot Age$

e.g. For an individual 50 years of age

$E(CD \mid Age = 50) = -.54 + .02 \cdot 50 = .46??$

Smooth Regression Estimate?



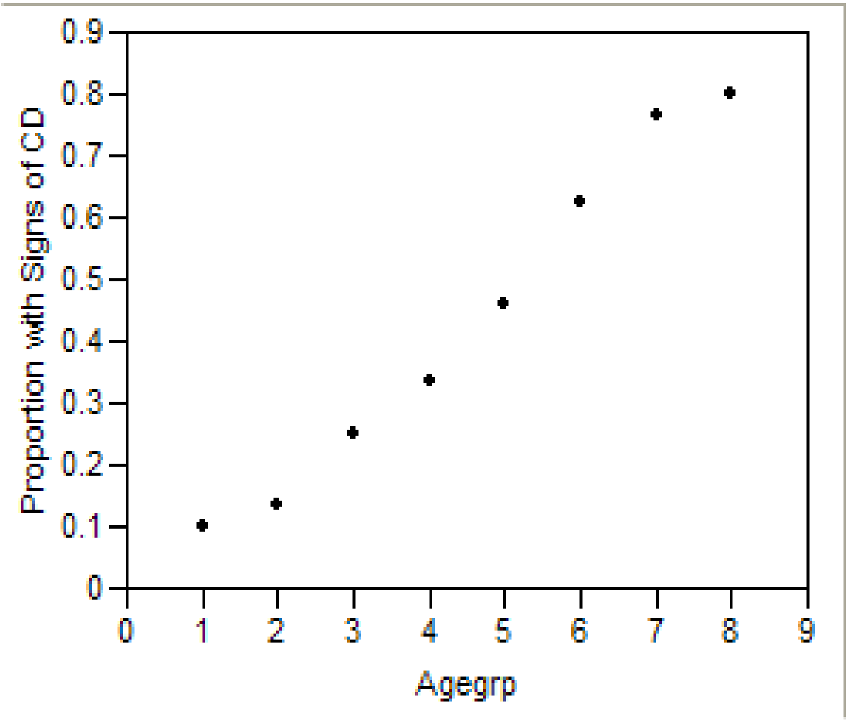
The smooth regression estimate is "S-shaped" but what does the estimated mean value represent?

Answer: P(CD|Age)!!!!

Logistic Regression

We can group individuals into age classes and look at the percentage/proportion showing signs of coronary heart disease.

Age group	# in group	Diseased #	Diseased Proportion
1) 20 - 29	10	1	.100
2) 30 - 34	15	2	.133
3) 35 - 39	12	3	.250
4) 40 - 44	15	5	.333
5) 45 - 49	13	6	.462
6) 50 - 54	8	5	.625
7) 55 - 59	17	13	.765
8) 60 – 64	10	8	.800



Notice the "S-shape" to the estimated proportions vs. age.

Logistic Function

$$P(\text{"}Success\text{"} \mid X) = \frac{e^{\beta_o + \beta_1 X}}{1 + e^{\beta_o + \beta_1 X}}$$

Logit Transformation

The logistic regression model is given by

$$P(Y \mid X) = \frac{e^{\beta_o + \beta_1 X}}{1 + e^{\beta_o + \beta_1 X}}$$

which is equivalent to

$$\underbrace{\ln\left(\frac{P(Y \mid X)}{1 - P(Y \mid X)}\right)} = \beta_o + \beta_1 X$$

This is called the
Logit Transformation

Dichotomous Predictor

Consider a dichotomous predictor (X) which represents the presence of risk (1 = present)

Disease (Y)	Risk Factor (X) Present (X = 1)	 Absent (X = 0)
Yes (Y = 1)	$P(Y=1 \mid X=1)$	$P(Y=1 \mid X=0)$
No (Y = 0)	$1-P(Y=1 \mid X=1)$	$1-P(Y=1 \mid X=0)$

$$\frac{P}{1-P}=e^{\beta_o+\beta_1 X}$$

$$\text{Odds for Disease with Risk Present} = \frac{\mathrm{P}(\mathrm{Y}=1 \mid \mathrm{X}=1)}{1-\mathrm{P}(\mathrm{Y}=1 \mid \mathrm{X}=1)} = e^{\beta_o+\beta_1}$$

$$\text{Odds for Disease with Risk Absent} = \frac{\mathrm{P}(\mathrm{Y}=1 \mid \mathrm{X}=0)}{1-\mathrm{P}(\mathrm{Y}=1 \mid \mathrm{X}=0)} = e^{\beta_o}$$

Therefore the odds ratio (OR)

$$=\frac{\text{Odds for Disease with Risk Present}}{\text{Odds for Disease with Risk Absent}}=\frac{e^{\beta_o+\beta_1}}{e^{\beta_o}}=e^{\beta_1}$$

Dichotomous Predictor

- Therefore, for the odds ratio associated with risk presence we have $OR = e^{\beta_1}$

- Taking the natural logarithm we have

$$\ln(OR) = \beta_1$$

Thus, the estimated regression coefficient associated with a 0-1 coded dichotomous predictor is the natural log of the OR associated with risk presence!!!

Logit is Directly Related to Odds

The logistic model can be written

$$\ln\left(\frac{P(Y \mid X)}{1-P(Y \mid X)}\right) = \ln\left(\frac{P}{1-P}\right) = \beta_o + \beta_1 X$$

This implies that the odds for success can be expressed as

$$\frac{P}{1-P} = e^{\beta_o + \beta_1 X}$$

This relationship is the key to interpreting the coefficients in a logistic regression model !!

Dichotomous Predictor (+1/-1 coding)

Consider a dichotomous predictor (X) which represents the presence of risk (1 = present)

	Risk Factor (X)	
Disease (Y)	**Present** (X = 1)	**Absent** (X = -1)
Yes (Y = 1)	$P(Y=1 \mid X=1)$	$P(Y=1 \mid X=-1)$
No (Y = 0)	$1-P(Y=1 \mid X=1)$	$1-P(Y=1 \mid X=-1)$

$$\frac{P}{1-P} = e^{\beta_o + \beta_1 X}$$

$$\text{Odds for Disease with Risk Present} = \frac{\text{P(Y = 1 | X = 1)}}{\text{1 - P(Y = 1 | X = 1)}} = e^{\beta_o + \beta_1}$$

$$\text{Odds for Disease with Risk Absent} = \frac{\text{P(Y = 1 | X = -1)}}{\text{1 - P(Y = 1 | X = -1)}} = e^{\beta_o - \beta_1}$$

Therefore the odds ratio (OR)

$$= \frac{\text{Odds for Disease with Risk Present}}{\text{Odds for Disease with Risk Absent}} = \frac{e^{\beta_o + \beta_1}}{e^{\beta_o - \beta_1}} = e^{2\beta_1}$$

Dichotomous Predictor

- Therefore, for the odds ratio associated with risk presence we have $OR = e^{2\beta_1}$

- Taking the natural logarithm we have

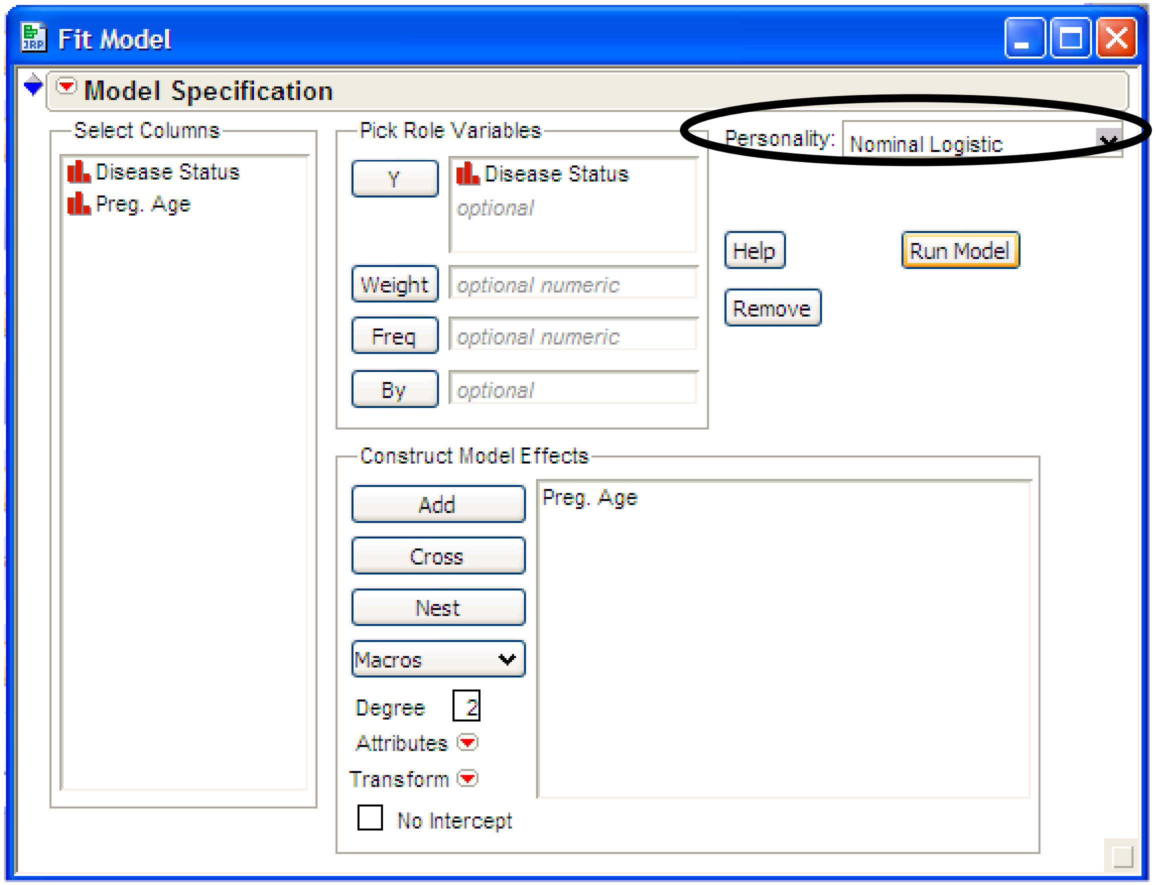
$$\ln(OR) = 2\beta_1$$

Thus, twice the estimated regression coefficient associated with a +1 / -1 coded dichotomous predictor is the natural log of the OR associated with risk presence!!!

Example: Age at 1st Pregnancy & Cervical Cancer

Use Fit Model   Y = Disease Status

X = Risk Factor Status



When the response Y is a dichotomous categorical variable the Personality box will automatically change to Nominal Logistic, i.e. Logistic Regression will be used.

Remember when a dichotomous categorical predictor is used JMP uses +1/-1 coding. If you want you can code them as 0-1 and treat is as numeric.

Example: Age at 1st Pregnancy & Cervical Cancer

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-2.1829122	0.2123468	105.68	<.0001*
Preg. Age[<= 25]	0.60737587	0.2123468	8.18	0.0042*

For log odds of Cervical/Control

$$\hat{\beta}_o = -2.183$$
$$\hat{\beta}_1 = 0.607$$

Thus the estimated odds ratio is

$$\ln(OR) = 2\hat{\beta}_1 = 2(.607) = 1.214$$
$$OR = e^{1.214} = 3.37$$

Women whose first pregnancy is at or before age 25 have 3.37 times the odds for developing cervical cancer than women whose 1st pregnancy occurs after age 25.

Example: Age at 1st Pregnancy & Cervical Cancer

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-2.1829122	0.2123468	105.68	<.0001*
Preg. Age[<= 25]	0.60737587	0.2123468	8.18	0.0042*

For log odds of Cervical/Control

$$\hat{\beta}_o = -2.183$$

$$\hat{\beta}_1 = 0.607$$

Thus the estimated odds ratio is

Odds Ratios

For Disease Status odds of Cervical versus Control

Odds Ratios for Preg. Age

Level1	/Level2	Odds Ratio	Reciprocal
> 25	<= 25	0.2967837	3.3694575

Risk Present → Odds Ratio for disease associated with risk presence

Example 1: Smoking and Low Birth Weight

Use Fit Model Y = Low Birth Weight (Low, Norm)

X = Smoking Status (Cig, NoCig)

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-2.0608189	0.0127482	26133	0.0000*
Smoking Status[Cig]	0.33493469	0.0127482	690.28	<.0001*

For log odds of Low/Norm

$$\hat{\beta}_1 = .335$$

$$OR = e^{2\hat{\beta}_1} = e^{.670} = 1.954$$

Odds Ratios

For Low Birth odds of Low versus Norm

Odds Ratios for Smoking Status

Level1	/Level2	Odds Ratio	Reciprocal
NoCig	Cig	0.5117754	1.9539821

We estimate that women who smoker during pregnancy have 1.95 times higher odds for having a child with low birth weight than women who do not smoke cigarettes during pregnancy.

Example 1: Smoking and Low Birth Weight

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-2.0608189	0.0127482	26133	0.0000*
Smoking Status[Cig]	0.33493469	0.0127482	690.28	<.0001*

For log odds of Low/Norm

$$\hat{\beta}_1 = .335$$

$$OR = e^{2\hat{\beta}_1} = e^{.670} = 1.954$$

Find a 95% CI for OR

1st Find a 95% CI for b_1

$$\hat{\beta}_1 \pm 1.96 SE(\hat{\beta}_1) = .335 \pm 1.96 \cdot (.013) = .335 \pm .025 = \underbrace{(.310,.360)}_{(LCL,UCL)}$$

2nd Compute CI for OR = (e^{2LCL}, e^{2UCL})

$$(e^{2\times 310}, e^{2\times .360}) = (1.86\,, 2.05)$$

We estimate that the odds for having a low birth weight infant are between 1.86 and 2.05 times higher for smokers than non-smokers, with 95% confidence.

Example 1: Smoking and Low Birth Weight

We might want to adjust for other potential confounding factors in our analysis of the risk associated with smoking during pregnancy. This is accomplished by simply adding these covariates to our model.

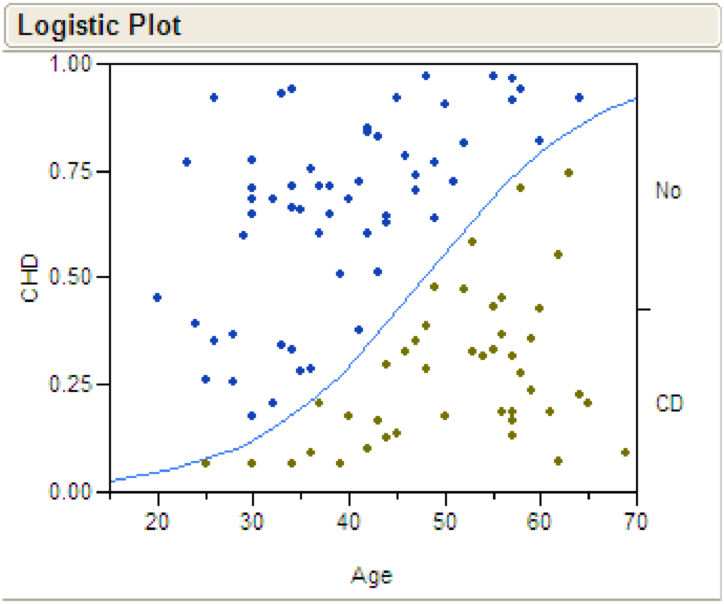
Multiple Logistic Regression Model

$$\ln\left(\frac{p}{1-p}\right) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Before looking at some multiple logistic regression examples we need to look at how continuous predictors and categorical variables with 3 or levels are handled in these models and how associated OR's are calculated.

Example 2: Signs of CD and Age

Fit Model Y = CD (CD if signs present, No otherwise)
X = Age (years)



Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-5.309453	1.1336546	21.94	<.0001*
Age	0.11092114	0.0240598	21.25	<.0001*

For log odds of CD/No

$$\ln\left(\frac{p}{1-p}\right) = -5.31 + .111 \cdot Age$$

$$\text{Odds} = \frac{p}{1-p} = e^{-5.31+.111Age}$$

Consider the risk associated with a ***c*** year increase in age.

$$\text{Odds Ratio (OR)} = \frac{\text{Odds for Age} = \text{x} + \text{c}}{\text{Odds for Age} = \text{x}} = \frac{e^{\beta_o+\beta_1(x+c)}}{e^{\beta_o+\beta_1 x}} = e^{c\beta_1}$$

Example 2: Signs of CD and Age

For example consider a 10 year increase in age, find the associated OR for showing signs of CD, i.e. $c = 10$

$$\textbf{OR} = e^{cb} = e^{10*.111} = 3.03$$

Thus we estimate that the odds for exhibiting signs of CD increase threefold for each 10 years of age. Similar calculations could be done for other increments as well.

For example for a $c = 1$ year increase

$OR = e^{b} = e^{.111} = 1.18$ or an 18% increase in odds per year

Example 2: Signs of CD and Age

- Can we assume that the increase in risk associated with a c unit increase is constant throughout one's life?
- Is the increase going from 20 → 30 years of age the same as going from 50 → 60 years?
- If that assumption is not reasonable then one must be careful when discussing risk associated with a continuous predictor.

Example 3: Race and Low Birth Weight

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-2.1979794	0.0165809	17572	0.0000*
Race[Black]	0.41029325	0.0190908	461.89	<.0001*
Race[Other]	-0.0890288	0.030963	8.27	0.0040*

For log odds of Low/Norm

$$Race[Black] = \begin{cases} +1 & \text{for race} = \text{black} \\ -1 & \text{for race} = \text{white} \end{cases}$$

$$Race[Other] = \begin{cases} +1 & \text{for race} = \text{other} \\ -1 & \text{for race} = \text{white} \end{cases}$$

Calculate the odds for low birth weight for each race (Low, Norm)

White Infants (reference group, missing in parameters)

$$e^{-2.198+.410(-1)-.089(-1)} = e^{-2.198-.410+.089} = .0805$$

Black Infants

$$e^{-2.198+.410(+1)-.089(0)} = .167$$

Other Infants

$$e^{-2.198+.410(0)-.089(+1)} = .102$$

OR for Blacks vs. Whites

= .167/.0805 = 2.075

OR for Others vs. Whites

= .102/.0805 = 1.267

OR for Black vs. Others

= .167/.102 = 1.637

Example 3: Race and Low Birth Weight

Finding these directly using the estimated parameters is cumbersome. JMP will compute the Odds Ratio for each possible comparison and their reciprocals in case those are of interest as well.

Odds Ratios

For Low Birth odds of Low versus Norm

Odds Ratios for Race

Level1	/Level2	Odds Ratio	Reciprocal
Other	Black	0.606942	1.6476038
White	Black	0.4811589	2.0783155
White	Other	0.7927592	1.261417

Odds Ratio column is odds for Low for Level 1 vs. Level 2.

Reciprocal is odds for Low for Level 2 vs. Level 1.   These are the easiest to interpret here as they represent increased risk.

Putting it all together

Now that we have seen how to interpret each of the variable types in a logistic model we can consider multiple logistic regression models with all these variable types included in the model.

We can then look at risk associated with certain factors adjusted for the other covariates included in the model.

Example 3: Smoking and Low Birth Weight

- Consider again the risk associated with smoking but this time adjusting for the potential confounding effects of education level and age of the mother & father, race of the child, total number of prior pregnancies, number children born alive that are now dead, and gestational age of the infant.

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	24.3444117	0.2557917	9057.9	0.0000*
Gender of child[1]	-0.1858494	0.01419	171.54	<.0001*
Age of father	-0.0012934	0.0030945	0.17	0.6760
Age of mother	0.00221874	0.003829	0.34	0.5623
Education of father (years)	0.00367121	0.0073201	0.25	0.6160
Education of mother (years)	-0.0047079	0.0074148	0.40	0.5255
Total Preg	-0.0444083	0.010651	17.38	<.0001*
BDead	0.15801032	0.0882007	3.21	0.0732
Smoker[Cigs]	0.38427179	0.0207736	342.18	<.0001*
Race[Black]	0.20104655	0.0289675	48.17	<.0001*
Race[Other]	0.07387835	0.0423638	3.04	0.0812
Gest Age	-0.7030361	0.0065632	11474	0.0000*

For log odds of Low/Norm

Several terms are not statistically significant and could consider using backwards elimination to simplify the model.

Example 3: Race and Low Birth Weight

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	22.8038037	0.2054094	12325	0.0000*
Gender of child[1]	-0.1784798	0.0122621	211.86	<.0001*
Total Preg	-0.0335729	0.0079688	17.75	<.0001*
BDead	0.1724931	0.0730957	5.57	0.0183*
Smoker[Cigs]	0.38081327	0.0163453	542.80	<.0001*
Race[Black]	0.21258524	0.0240098	78.40	<.0001*
Race[Other]	0.07032665	0.0378297	3.46	0.0630
Gest Age	-0.6610415	0.0054805	14548	0.0000*

For log odds of Low/Norm

None of the mother and farther related covariates entered into the final model.

Adjusting for the included covariates we find smoking is statistically significant ($p < .0001$)

Odds Ratios for Smoker

Level1	/Level2	Odds Ratio	Reciprocal
No	Cigs	0.4669064	2.141757

Adjusting for the included covariates we find the odds ratio for low birth weight associated with smoking during pregnancy is 2.142.

Odds Ratios for the other factors in the model can be computed as well. All of which can be prefaced by the "**adjusting for…**" statement.

Summary

- In logistic regression the response (Y) is a dichotomous categorical variable.
- The parameter estimates give the odds ratio associated the variables in the model.
- These odds ratios are adjusted for the other variables in the model.
- One can also calculate $P(Y|\mathbf{X})$ if that is of interest, e.g. given demographics of the mother what is the estimated probability of her having a child with low birth weight.

Interpretation of a single *categorical* parameter

- If your reference group is level 0, then the coefficient of β_k represents the difference in the **log odds** between level k of your variable and level 0.
- Therefore, $\boldsymbol{e^{\beta}}$ **is an odds ratio** for category *k* vs. the reference category of *x*.

Hypothesis testing

- Significance tests focuses on a test of H_0: $\beta = 0$ vs. H_a: $\beta \neq 0$.
- The Wald, Likelihood Ratio, and Score test are used (we'll focus on Wald method)
- Wald CI easily obtained, score and LR CI numerically obtained.
- For Wald, the 95% CI (on the log odds scale) is

$$\hat{\beta} \pm 1.96(SE(\hat{\beta}))$$

95% CI for parameter

- Similarly, the Wald 95% CI for the odds ratio is obtained by exponentiation.
- The following yields the lower and upper 95% confidence limits:

$$\exp(\hat{\beta} \pm 1.96(SE(\hat{\beta})))$$

- 1.96 corresponds to $z_{0.05/2}$, where $z \sim N(0,1)$

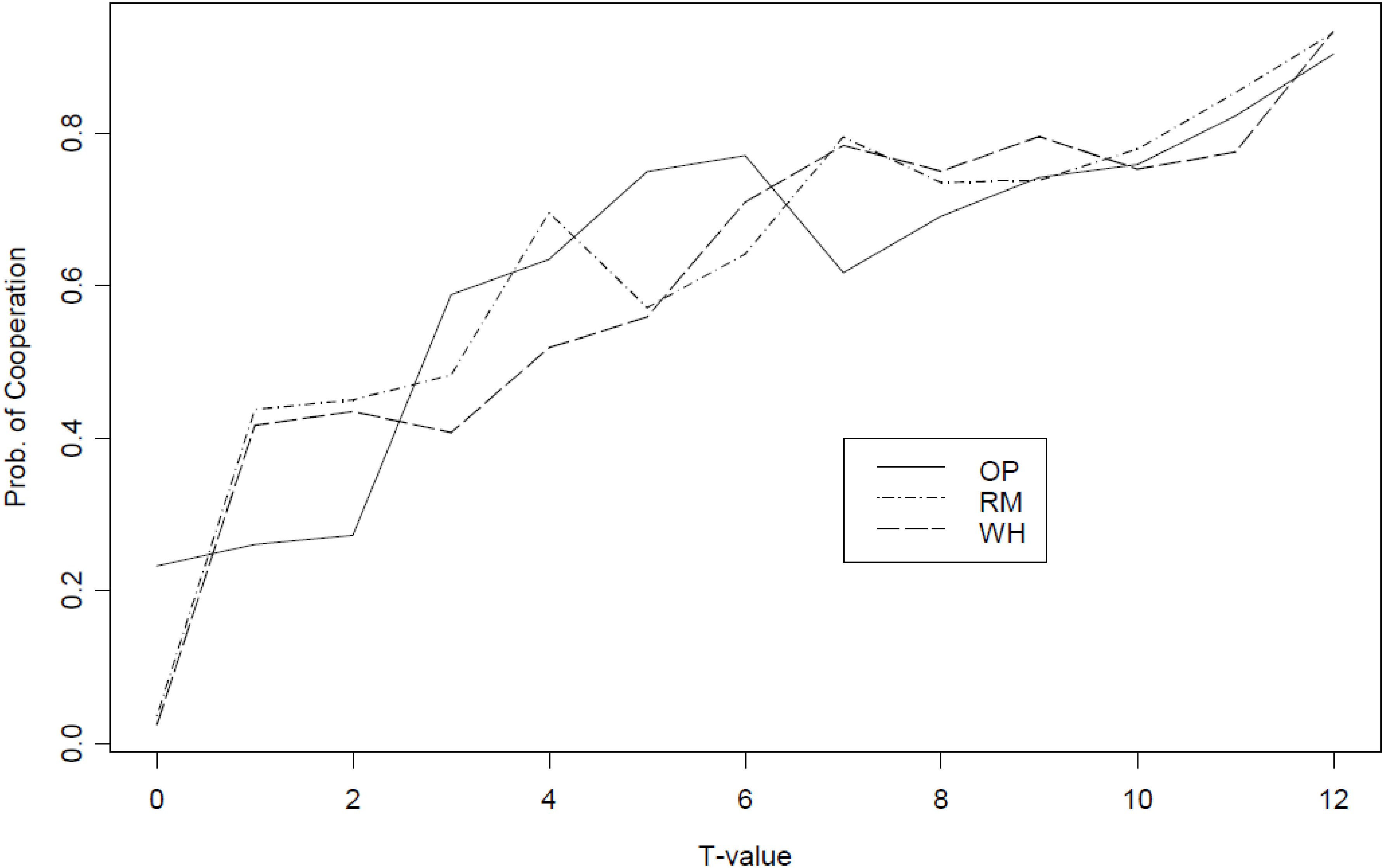
Hypothesis testing (ctd)

- The Wald statistic of the test H_0: $\beta = \beta_0$ is

$$\frac{(\hat{\beta} - \beta_0)^2}{\mathrm{var}(\hat{\beta})} \sim \chi_1^2$$

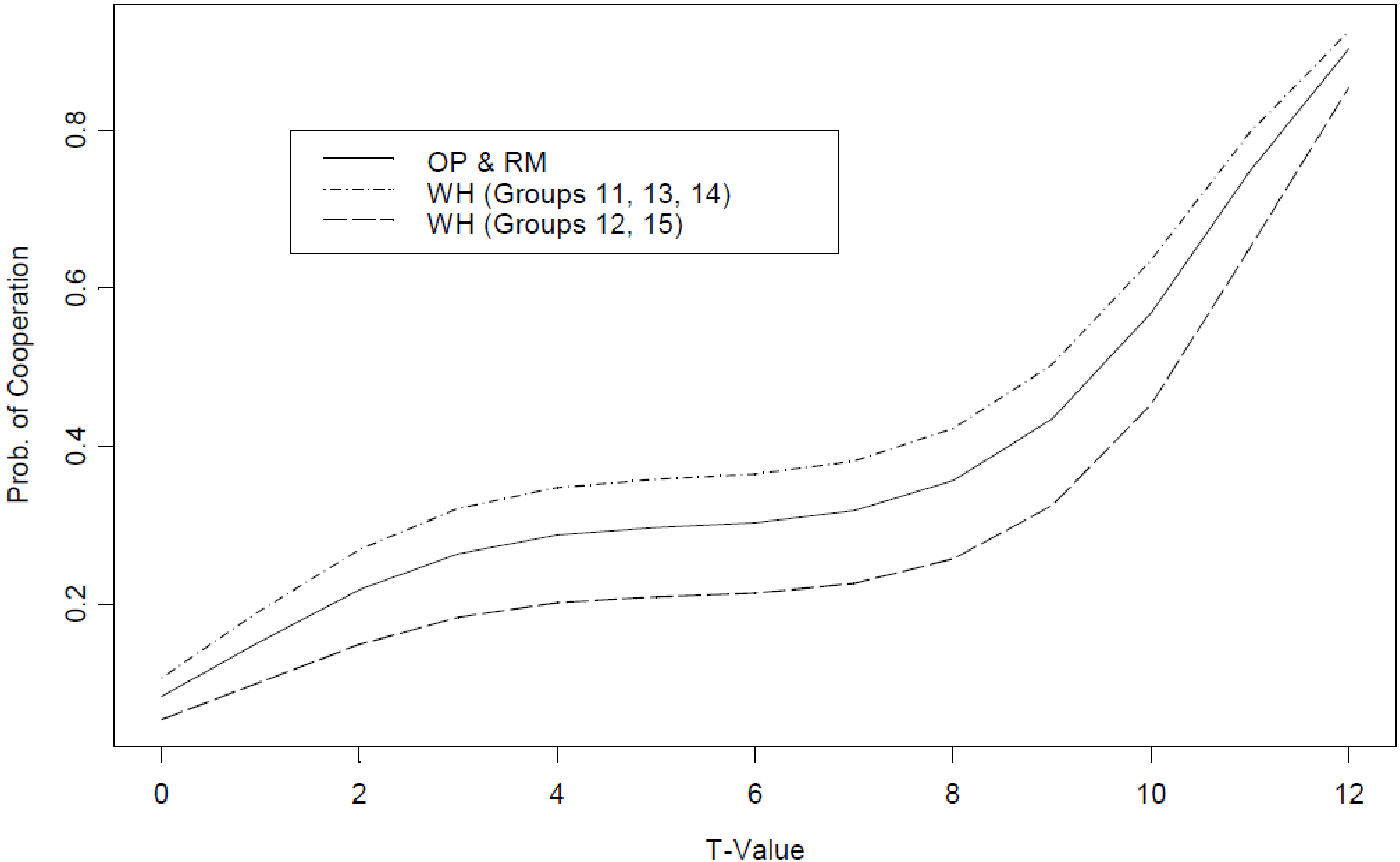
- Under H_0, the test statistic is asymptotically chi-sq. with 1 df (at $\alpha = 0.05$, the critical value is 3.84).

範例四、雙人賽局（囚犯困境）合作機率



$$\operatorname{logit}\left(p(C)\right) = \beta_0 + \beta_1 T + \beta_2 T^2 + \beta_3 T^3 + \text{other terms}$$

羅吉士迴歸估計結果（囚犯困境）



$$\text{logit}(p) = -2.3999 + .8277T - .1520T^2 + .0096T^3 + .2766\,Trt\,3 - .7458\,\text{Gr}12\&15$$

$$\quad(.0889)\quad(.0855)\quad(.0201)\quad(.0012)\quad(.1090)\quad(.1762)$$

羅吉士迴歸模型評估（囚犯困境）

	Model 1 (RM/WH)	*Model 2 (RM/WH)*	*Model 3 (RM/WH/WHP)*	*Model 4 (RM/WH/WHP)*
Intercept	-2.5279	-3.1827	-3.0029	-3.1000
T	0.8737	0.8762	0.6943	0.6893
T^2	-0.1643	-0.1592	-0.1115	-0.1113
T^3	0.0106	0.0102	0.0069	0.0069
Groups Dummy	-0.7035			
WH Treatment Dummy	0.3243 (.011)			
Avg. Payoff		0.1412	0.1633	0.1780
Information Dummy				0.2723 (.034)
Log (likelihood)	-992.616	-991.365	-1733.115	-1730.921
Goodness-of-Fit	.700	.244	.155	.248
Concordant	79.8%	80.6%	78.5%	78.8%

Note: All estimations are significant with $p < 0.0001$ unless noted otherwise in parentheses.

References

1. Paul D. Allison, “Logistic Regression Using the SAS System: Theory and Application”, SAS Institute, Cary, North Carolina, 1999.
2. Alan Agresti, “Categorical Data Analysis”, 2nd Ed., Wiley Interscience, 2002.
3. David W. Hosmer and Stanley Lemeshow “Applied Logistic Regression”, Wiley-Interscience, 2nd Edition, 2000.

累進機率與Logit

- 我們在此處用累進機率cumulative probabilities的概念作為基礎
- 令$P(y \leqq j)$代表回答落在j這個類屬或以下的機率（1, 2, …,j）
- Multinomial Logit 可逐項拆解：

→$P(y=1)$

→$P(y \leqq 2)=P(y=1)+P(y=2)$

→$P(y \leqq n)=\mathrm{P}(\mathrm{y}=1)+\ldots+\mathrm{P}(\mathrm{y}=n)$

累進機率與Logit

- 每個類屬 j 或以下的勝算odds是

  $$P(y \leq j) / P(y > j)$$

- 每一個累進機率都可以被轉換成「高於」或「低於」的二元變數的勝算
- A popular logistic model for an ordinal response uses logits of the cumulative probabilities

Cumulative logits

- 如果選項只有三種，則可得：

$$\log it[P(y \le 1)] = \log[\frac{P(y=1)}{P(y>1)}] = \log[\frac{P(y=1)}{P(y=2)+P(y=3)}]$$

$$\log it[P(y \le 2)] = \log[\frac{P(y \le 2)}{P(y>2)}] = \log[\frac{P(y=1)+P(y=2)}{P(y=3)}]$$

$$\log it[P(y \le 3)] = 1$$

Cumulative Logit Models for an Ordinal Response

- A model can simultaneously describe the effect of an explanatory variable on all the cumulative probabilities for *y*.
- 對於每個累積機率，這個模型就像是一般的羅吉斯模型，每一組自變項都可分成「高於」和「低於」特定的類屬*j*。
- 這個模型是

  $$\text{Logit}[P(y \leqq j)] = \alpha_j + \beta x,\ j = 1, 2, \ldots, c-1.$$

- In this model, β does not have a *j* subscript.
- It has the same value for each cumulative logit. In other words, the model assumes that the effect of *x* is the same for each cumulative probability.
- This cumulative logit model with this common effect is often called the proportional odds model(比例勝算模型).

Cumulative Logit Models for an Ordinal Response

- For each j, the odds that $y \leqq j$ multiply by e^{β} for each one-unit increase in x.
- Model fitting treats the observations as independent from a multinomial distribution.
- This is a generalization of the binomial distribution from two to multiple outcome categories.
- Software estimates the parameters using all the cumulative probabilities at once. This provides a single estimate beta-hat for the effect of x, rather than the separate estimates.