Introduction to Regression

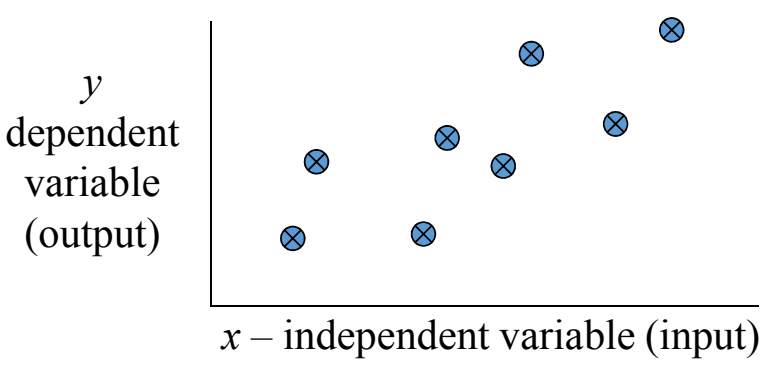
Linear & Logistic Regression

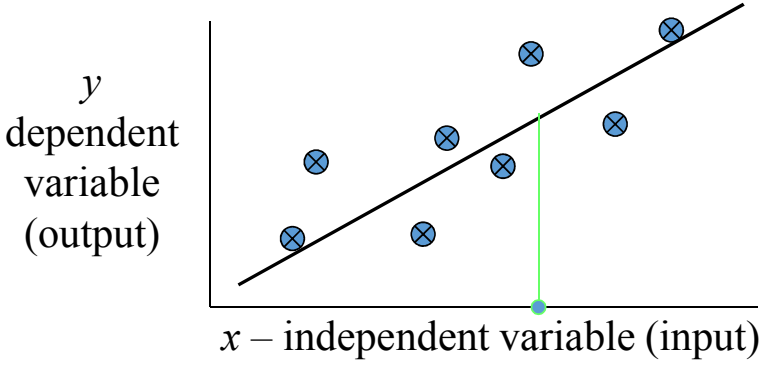
Multivariate analysis

- Multiple models
  - Linear regression
  - Logistic regression
  - Cox model
  - Poisson regression
  - Loglinear model
  - Discriminant analysis
  - Etc.

Regression

- Metode statistika yang digunakan untuk mengetahui hubungan antara variabel terikat (dependen;Y) dengan satu atau lebih variabel bebas (independen;X)
- A form of statistical modeling that attempts to evaluate the relationship between one variable (termed the dependent variable) and one or more other variables (termed the independent variables).
- It is a form of global analysis as it only produces a single equation for the relationship. Fit data with the best hyper-plane which "goes through" the points
- A model for predicting one variable from another.





Linear Regression

- Disebut juga dengan istilah *Ordinary Least Squares (OLS) regression.*
- Regression used to fit a linear model to data where the dependent variable is continuous:
- *Types:*
  - Simple linear regression
  - Multiple Linear Regression

Simple linear regression

- Assume just one (input) independent variable *x*, and one (output) dependent variable *y*
- $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

Multiple Linear Regression

- Assume a vector of input as independent variable $x_1, x_2, \ldots, x_n$, and one (output) dependent variable *y*
- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$

Kegunaan: Menghitung peluang

- Persamaan yang diperoleh dari proses regresi logistik, dapat digunakan untuk menghitung peluang responden diluar responden yang termasuk dalam penelitian.
- Contoh: Proses pengajuan kredit
  - Evaluasi kelayakan seseorang layak atau tidak untuk menerima kredit pinjaman dari bank.
  - Peluang calon penerima kredit tersebut untuk bisa mengembalikan pinjaman atau tidak, nilai antara 0 – 1.
- Menggunakan **data** calon peminjam untuk menentukan peluang :
  - memiliki pendapatan dibawah Rp. X
  - memiliki pinjaman yang telah dimiliki sebelumnya sejumlah Rp. X.
  - tanggungan kerja sebesar Y
- Data masukan ini disebut **Variabel Bebas (Independent Variable)**.

Simple linear regression

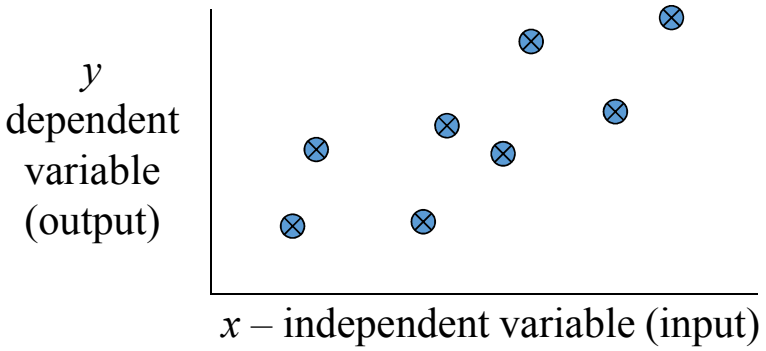
- Given a set of points (X_i, Y_i), we wish to find a linear function (or line in 2 dimensions) that "goes through" these points.
- In general, the points are not exactly aligned:

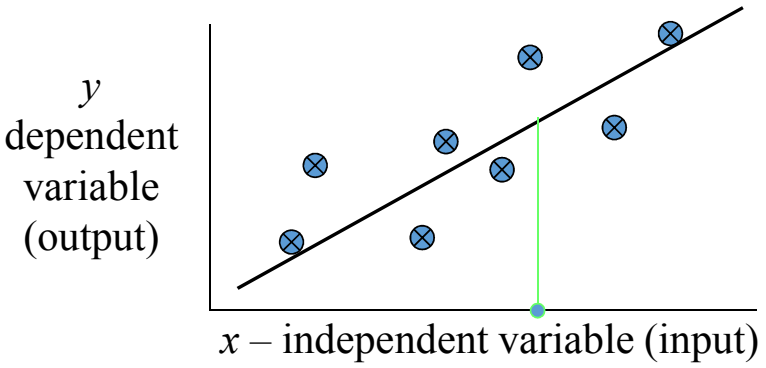
  Find line that best fits the points

- **Which line should we use?**
  - ❑ Choose an objective function
  - ❑ For simple linear regression we choose sum squared error (SSE)

    S ($predicted_i - actual_i$)^2 = S ($residue_i$)^2

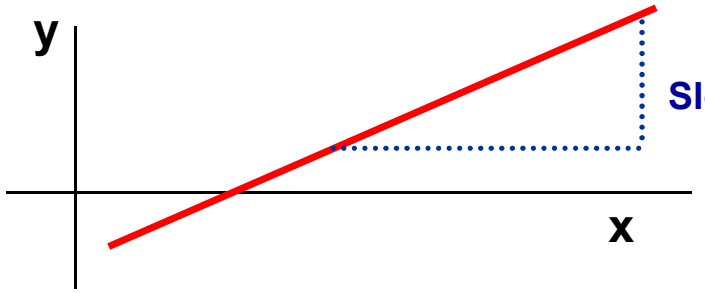
  - ❑ Thus, find the line which minimizes the sum of the squared residues (e.g. least squares)





Simple linear regression

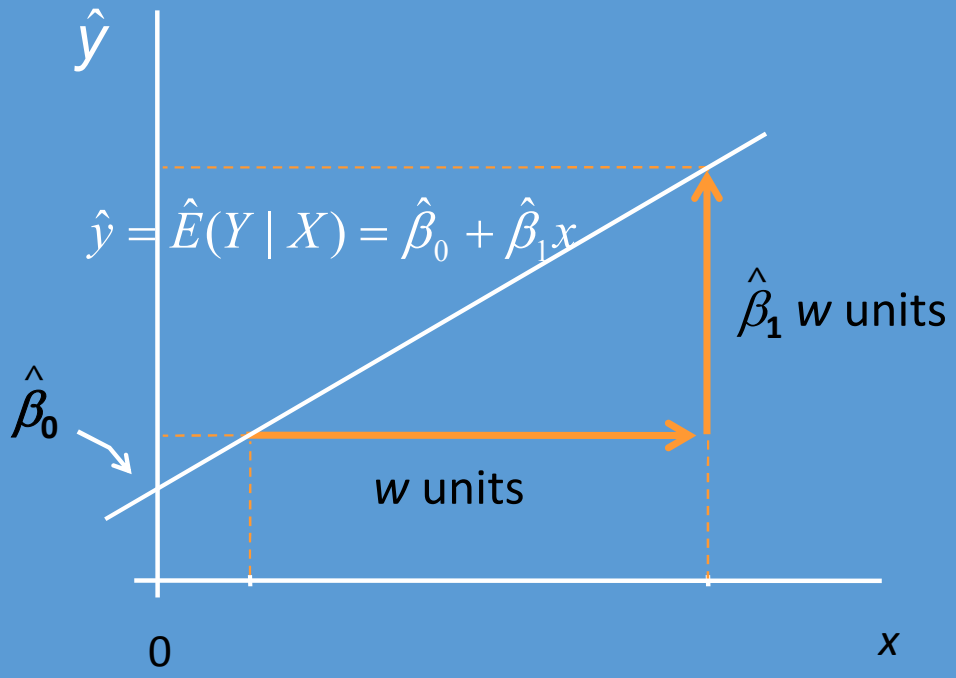
- Example: Relation between 2 continuous variables (systolic blood pressure (SBP) and age)



$$y = \alpha + \beta_1 x_1$$

- Regression coefficient β_1
  - Measures association between y and x
  - Amount by which y changes on average when x changes by one unit
  - Least squares method –> Sum Squared Error (SSE)

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!

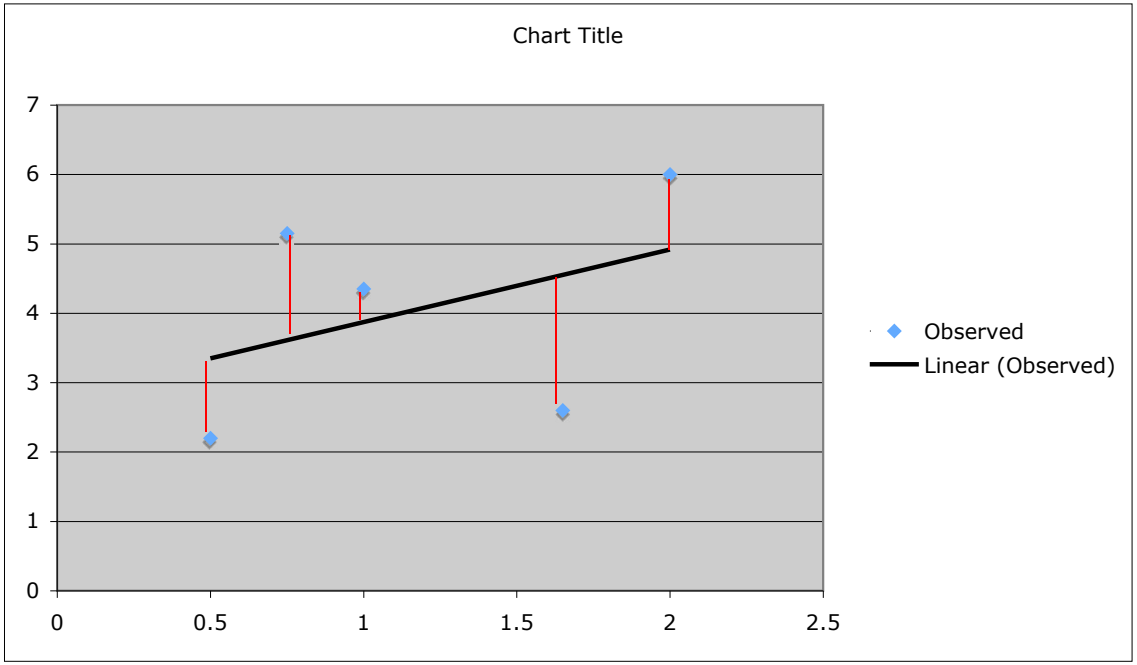
Multiple linear regression

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

Predicted	Predictor variables
Response variable	Explanatory variables
Outcome variable	Covariables
Dependent	Independent variables

Residue

- Error or residue:
  - Observed value - Predicted value



The coefficient of determination is used as a measure of how well a **regression** line explains the relationship between a dependent variable (Y) and an independent variable (X).

Sum-squared Error (SSE)

$$SSE = \sum_{y} (y_{observed} - y_{predicted})^2$$

Total sum of squares (TSS):

$$TSS = \sum_{y} (y_{observed} - \bar{y}_{observed})^2$$

$$R^2 = 1 - \frac{SSE}{TSS}$$

What is Best Fit?

- The smaller the SSE, the better the fit
- Hence,
  - Linear regression attempts to minimize SSE (or similarly to maximize R2)
- Assume 2 dimensions $Y = \beta_0 + \beta_1 X$

- To find the values for the coefficients which minimize the objective function we take the partial derivates of the objective function (SSE) with respect to the coefficients. Set these to 0, and solve.

$$\beta_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

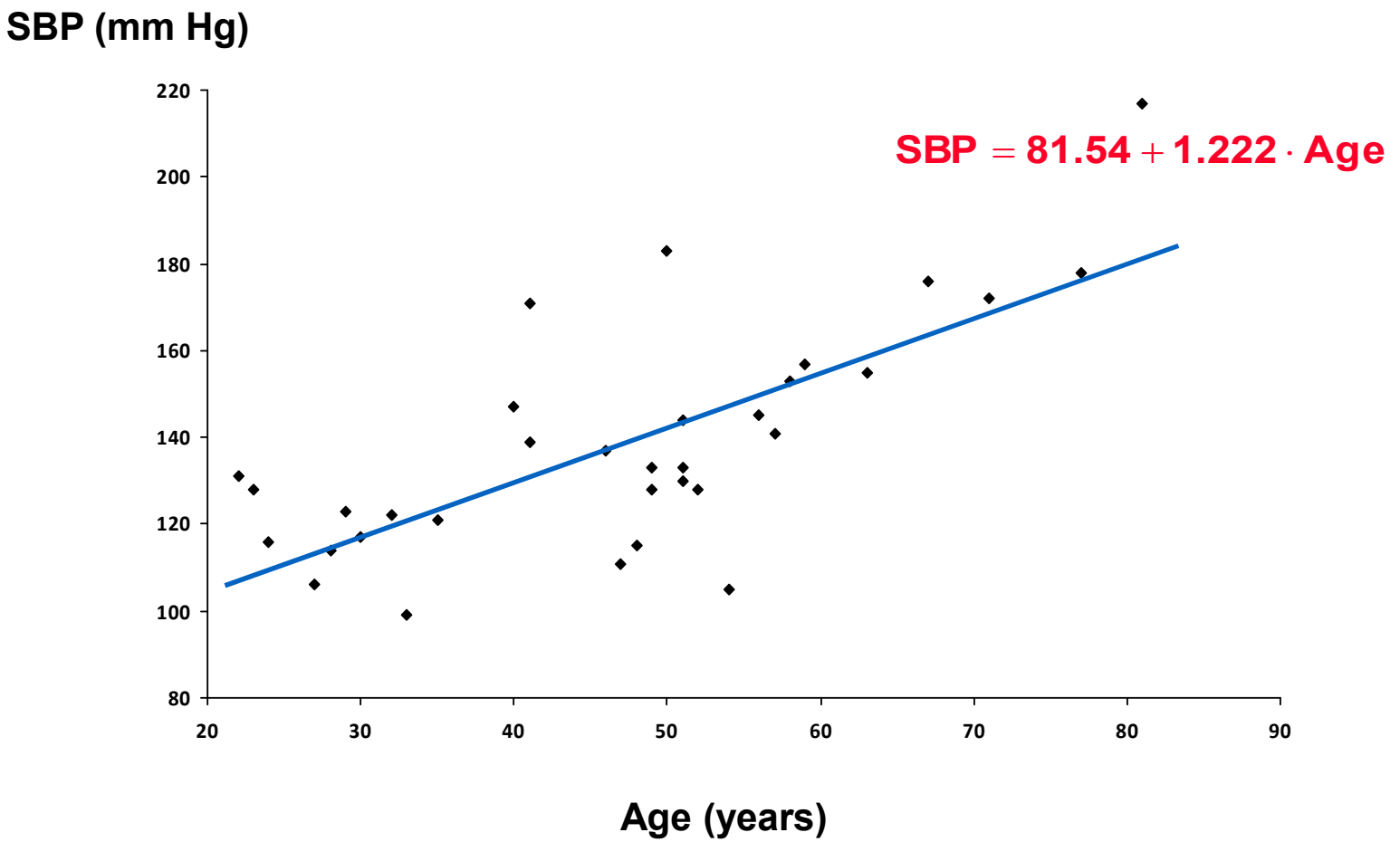
$$\beta_0 = \frac{\sum y - \beta_1 \sum x}{n}$$

Example I: Simple linear regression

Table 1 Age and systolic blood pressure (SBP) among 33 adult women

Age	SBP	Age	SBP	Age	SBP
22	131	41	139	52	128
23	128	41	171	54	105
24	116	46	137	56	145
27	106	47	111	57	141
28	114	48	115	58	153
29	123	49	133	59	157
30	117	49	128	63	155
32	122	50	183	67	176
33	99	51	130	71	172
35	121	51	133	77	178
40	147	51	144	81	217

Example I: Simple linear regression (cont.)



adapted from Colton T. Statistics in Medicine. Boston: Little Brown, 1974

Example II: linear regression

Coefficients computation

x	y	x^2	xy
1.20	4.00	1.44	4.80
2.30	5.60	5.29	12.88
3.10	7.90	9.61	24.49
3.40	8.00	11.56	27.20
4.00	10.10	16.00	40.40
4.60	10.40	21.16	47.84
5.50	12.00	30.25	66.00
24.10	**58.00**	**95.31**	**223.61**

$$\beta_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$= \frac{7 \times 223.61 - 24.10 \times 58.00}{7 \times 95.31 - 24.10^2}$$

$$= \frac{1565.27 - 1397.80}{667.17 - 580.81}$$

$$= \frac{167.47}{86.36} = \underline{\underline{1.94}}$$

$$\beta_0 = \frac{\sum y - \beta_1 \sum x}{n}$$

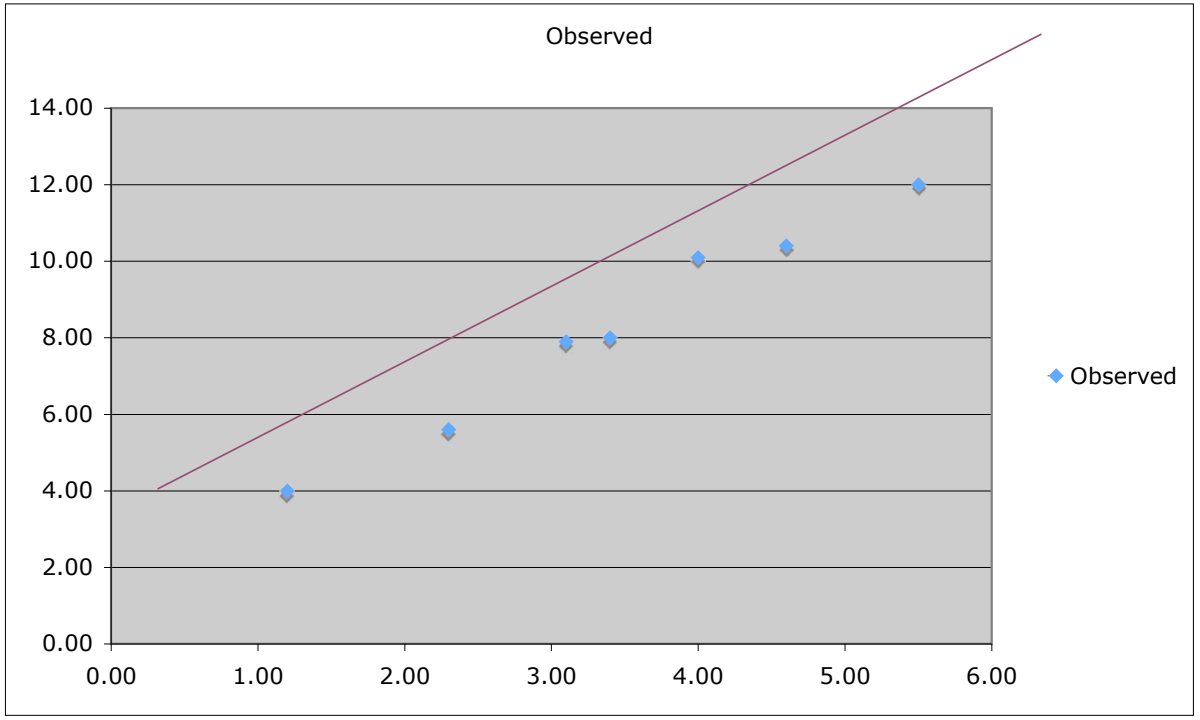
$$= \frac{58.00 - 1.94 \times 24.10}{7}$$

$$= \frac{11.27}{7} = \underline{\underline{1.61}}$$

y=1.94x+1.61

Example II: linear regression (cont.)

Drawing the lines



Example II: linear regression (cont.)

SEE and TSS computation

x	y (obs)	y (pred)	SSE	TSS
1.20	4.00	3.94	0.004	18.367
2.30	5.60	6.07	0.221	7.213
3.10	7.90	7.62	0.078	0.149
3.40	8.00	8.21	0.044	0.082
4.00	10.10	9.37	0.533	3.292
4.60	10.40	10.53	0.017	4.470
5.50	12.00	12.28	0.078	13.796
			0.975	**47.369**

$$R^2 = 1 - \frac{SSE}{TSS} = 1 - \frac{0.975}{47.369} = 0.98$$

Logistic Regression

- Regression used to fit a curve to data in which the dependent variable is binary, or dichotomous.
- Example: Medicine
  - We might want to predict response to treatment, where we might code survivors as 1 and those who don't survive as 0
- Logistic regression fits the data with a sigmoidal/logistic curve rather than a line and outputs an approximation of the probability of the output given the input

Logistic Regression

- Regresi logistik adalah sebuah pendekatan untuk membuat model prediksi seperti halnya regresi linear.
- Perbedaannya adalah pada regresi logistik, peneliti memprediksi variabel terikat yang berskala dikotomi.
- Skala dikotomi yang dimaksud adalah skala data nominal dengan dua kategori, misalnya: Ya dan Tidak, Baik dan Buruk atau Tinggi dan Rendah.

Example: Logistic regression (1)

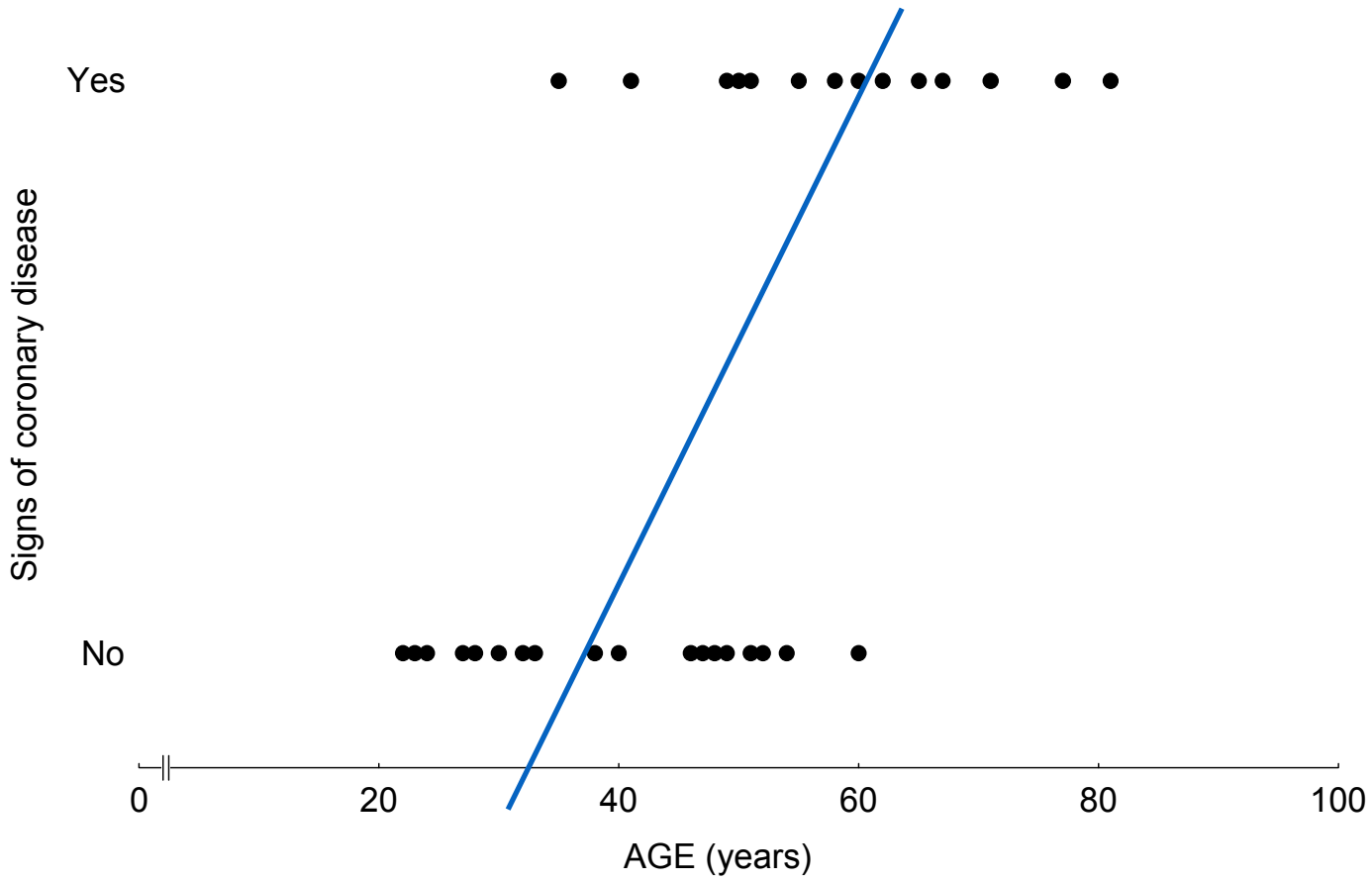
Table 2 Age and signs of coronary heart disease (CD)

Age	CD	Age	CD	Age	CD
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

How can we analyse these data?

- Compare mean age of diseased and non-diseased
  - Non-diseased: 38.6 years
  - Diseased: 58.7 years ($p<0.0001$)
- Linear regression?

Dot-plot: Data from Table 2



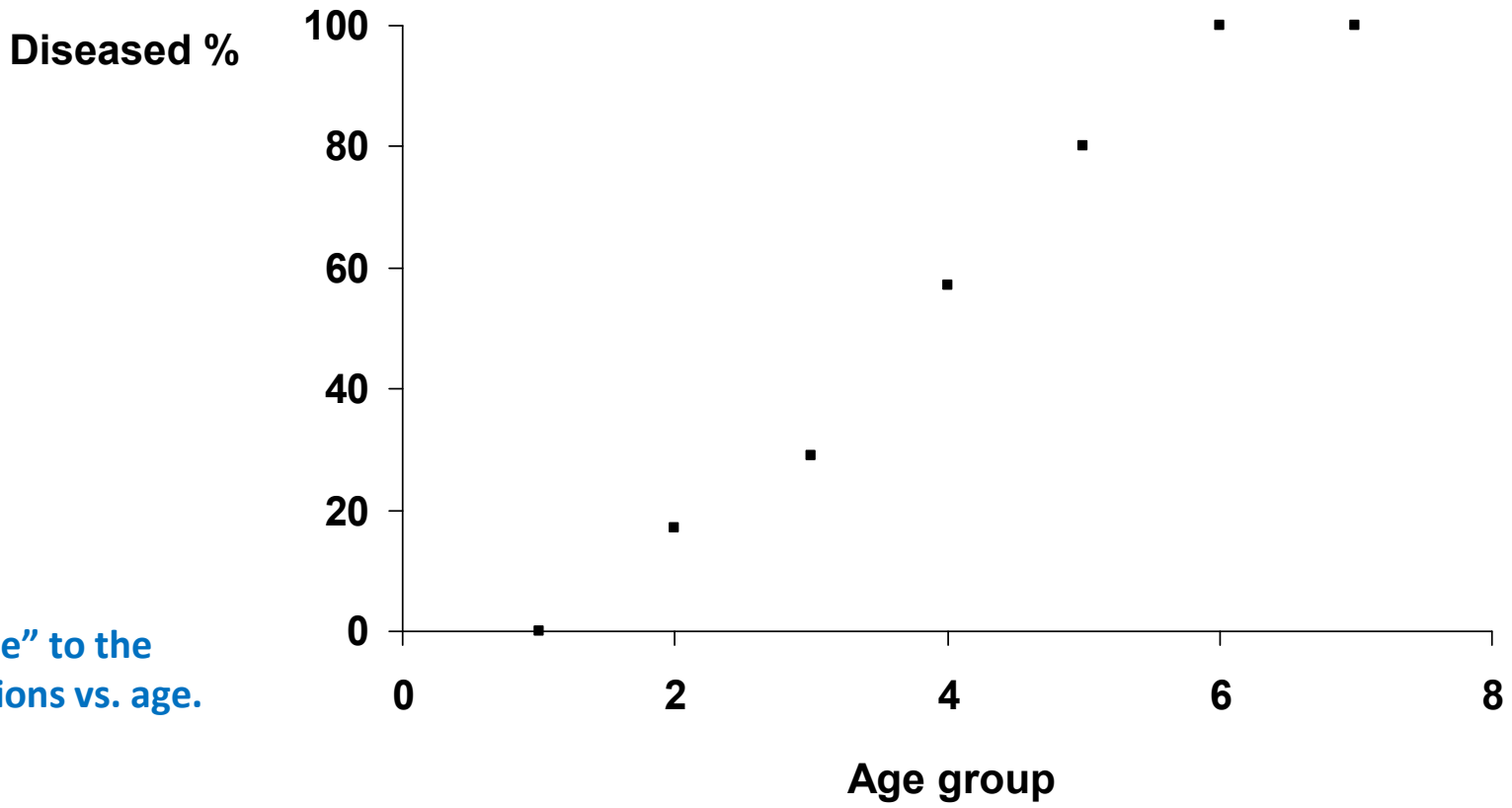
Logistic regression

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

Table 3 Prevalence (%) of signs of CD according to age group

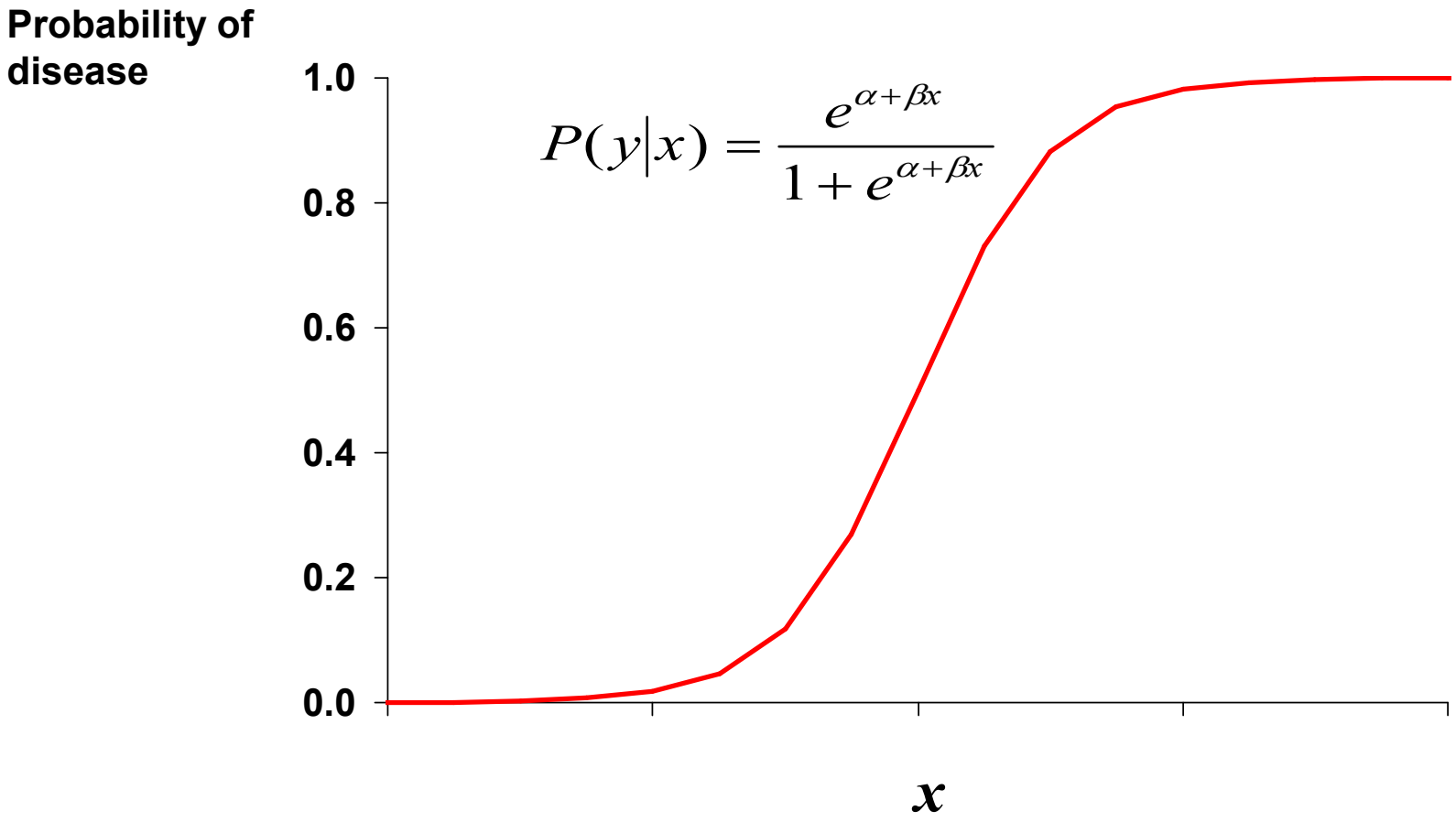
		Diseased	
Age group	# in group	#	%
20 - 29	5	0	0
30 - 39	6	1	17
40 - 49	7	2	29
50 - 59	7	4	57
60 - 69	5	4	80
70 - 79	2	2	100
80 - 89	1	1	100

Dot-plot: Data from Table 3



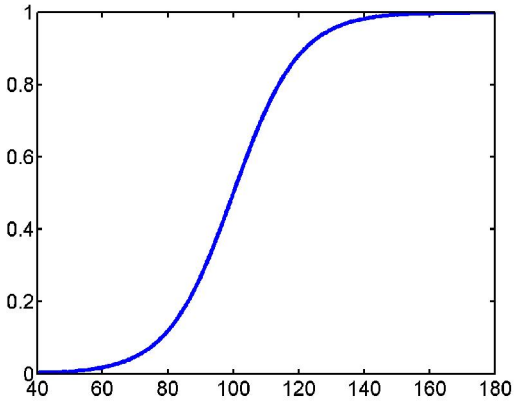
Notice the “S-shape” to the estimated proportions vs. age.

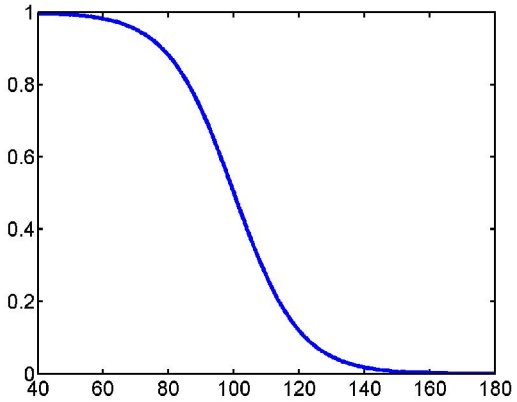
Logistic function (1)



Logistic Response Function

- When the response variable is binary, the shape of the response function is often sigmoidal:





Example

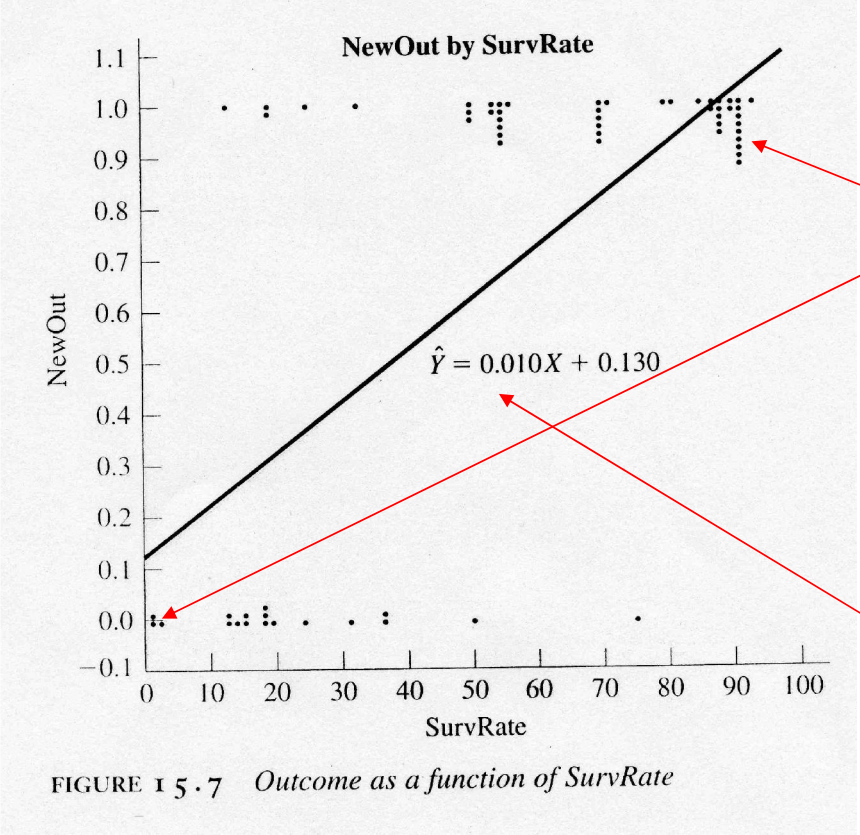


FIGURE 15.7 *Outcome as a function of SurvRate*

Observations:
For each value of SurvRate, the number of dots is the number of patients with that value of NewOut

Regression:
Standard linear regression

Problem: extending the regression line a few units left or right along the X axis produces predicted probabilities that fall outside of [0,1]

A Better Solution

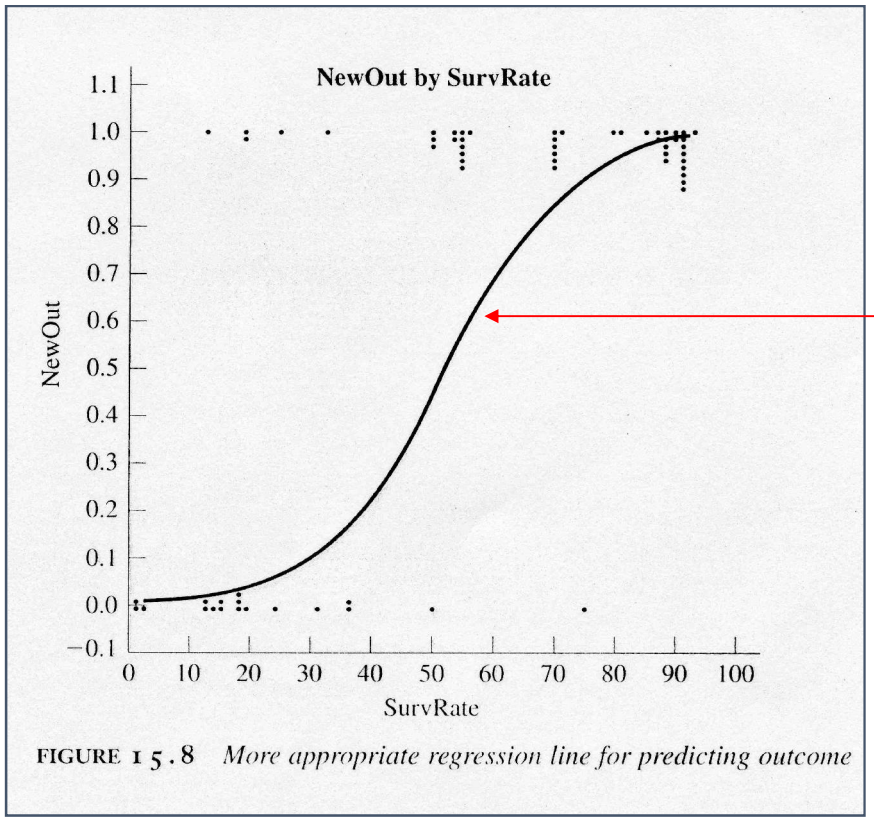


FIGURE 15.8 *More appropriate regression line for predicting outcome*

Regression Curve:
Sigmoid function!

(bounded by asymptotes $y=0$ and $y=1$)

Logit Transformation

The logistic regression model is given by

$$P(y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \quad \sim \quad \frac{P(y|x)}{1-P(y|x)}$$

which is equivalent to

$$\ln\left[\frac{P(y|x)}{1-P(y|x)}\right] = \alpha + \beta x$$

logit of *P(y|x)*

This is called the Logit Transformation

- ✓α = **log odds of disease in unexposed**
- ✓β = **log odds ratio associated with being exposed**
- ✓e^{β} = **odds ratio**

Dichotomous Predictor

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

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

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

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

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

Therefore the odds ratio (OR) $= \dfrac{\text{Odds for Disease with Risk Present}}{\text{Odds for Disease with Risk Absent}} = \dfrac{e^{\beta_o+\beta_1}}{e^{\beta_o}} = e^{\beta_1}$

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)	$P(Y=1\|X=1)$	$P(Y=1\|X=-1)$
No (Y = 0)	$1-P(Y=1\|X=1)$	$1-P(Y=1\|X=-1)$

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

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

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

Therefore the odds ratio (OR) $= \dfrac{\text{Odds for Disease with Risk Present}}{\text{Odds for Disease with Risk Absent}} = \dfrac{e^{\beta_o+\beta_1}}{e^{\beta_o-\beta_1}} = e^{2\beta_1}$

Dichotomous Predictor

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

- Taking the natural logarithm we have

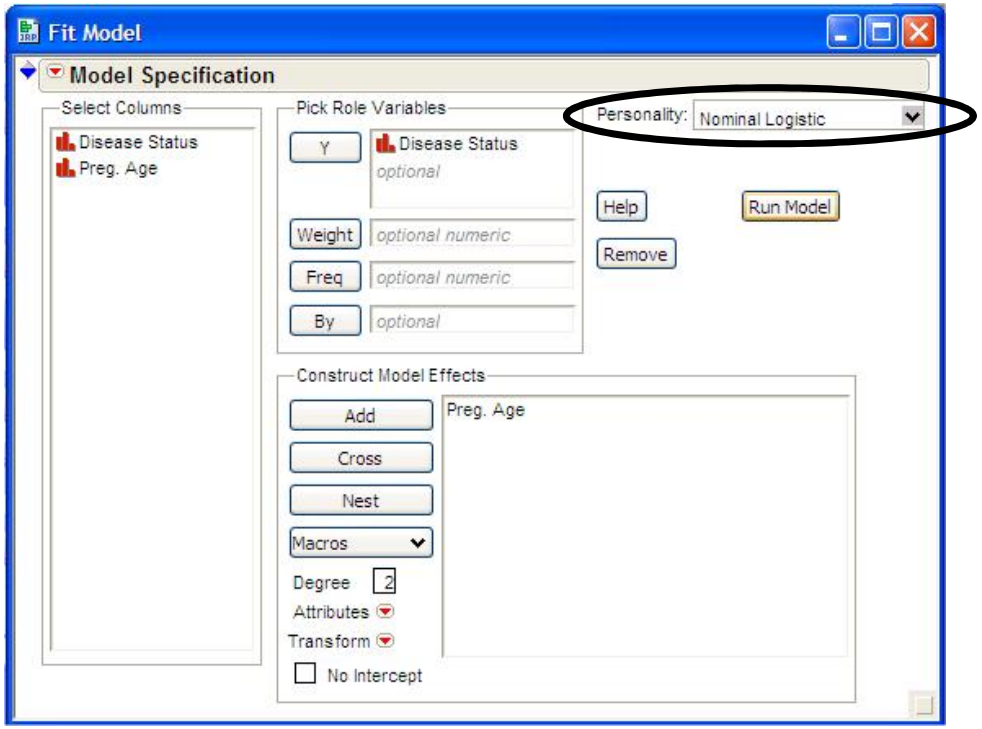
$$\ln(OR) = 2\beta_1$$

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 and 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 and 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 and 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

Fitting equation to the data

- Linear regression: Least squares
- Logistic regression: Maximum likelihood
- Likelihood function
  - Estimates parameters α and β
  - Practically easier to work with log-likelihood

$$L(\mathrm{B}) = \ln\left[l(\mathrm{B})\right] = \sum_{i=1}^{n} \left\{ y_i \ln\left[\pi(x_i)\right] + (1 - y_i) \ln\left[1 - \pi(x_i)\right] \right\}$$

Least squares

- Disebut juga metode pendugaan kuadrat terkecil
- Adalah salah satu metode yang sering digunakan untuk mendapatkan nilai-nilai penduga dalam pemodelan regresi yang meminimumkan jumlah kuadrat galat (Sum of Square Error).

Maximum likelihood

- Disebut juga metode kemungkinan maksimum
- Adalah salah satu metode pendugaan yang memaksimalkan fungsi likelihood $L(\theta|Y)$.

Dasar pemikiran Maximum likelihood

- Misal terdapat sebuah kotak yang memuat 3 bola.
- Diketahui bahwa setiap bola MUNGKIN berwarna merah atau putih, tetapi tidak diketahui banyaknya bola untuk setiap warna.
- Dipilih sample secara random 2 buah boleh tanpa pengembalian.
- Jika sample random menghasilkan 2 bola merah, dapat disimpulkan bahwa jumlah bola merah pada kotak haruslah 2 atau 3.
- (Jika terdapat 0 atau 1 bola merah pada kotak, maka tidak mungkin untuk memperoleh 2 bola merah ketika mengambil sampel tanpa pengembalian).
- Jika terdapat 2 bola merah dan 1 bola putih pada kotak, peluang terpilihnya 2 bola merah secara acak adalah

$$\frac{\binom{2}{2}\binom{1}{0}}{\binom{3}{2}} = \frac{1}{3}$$

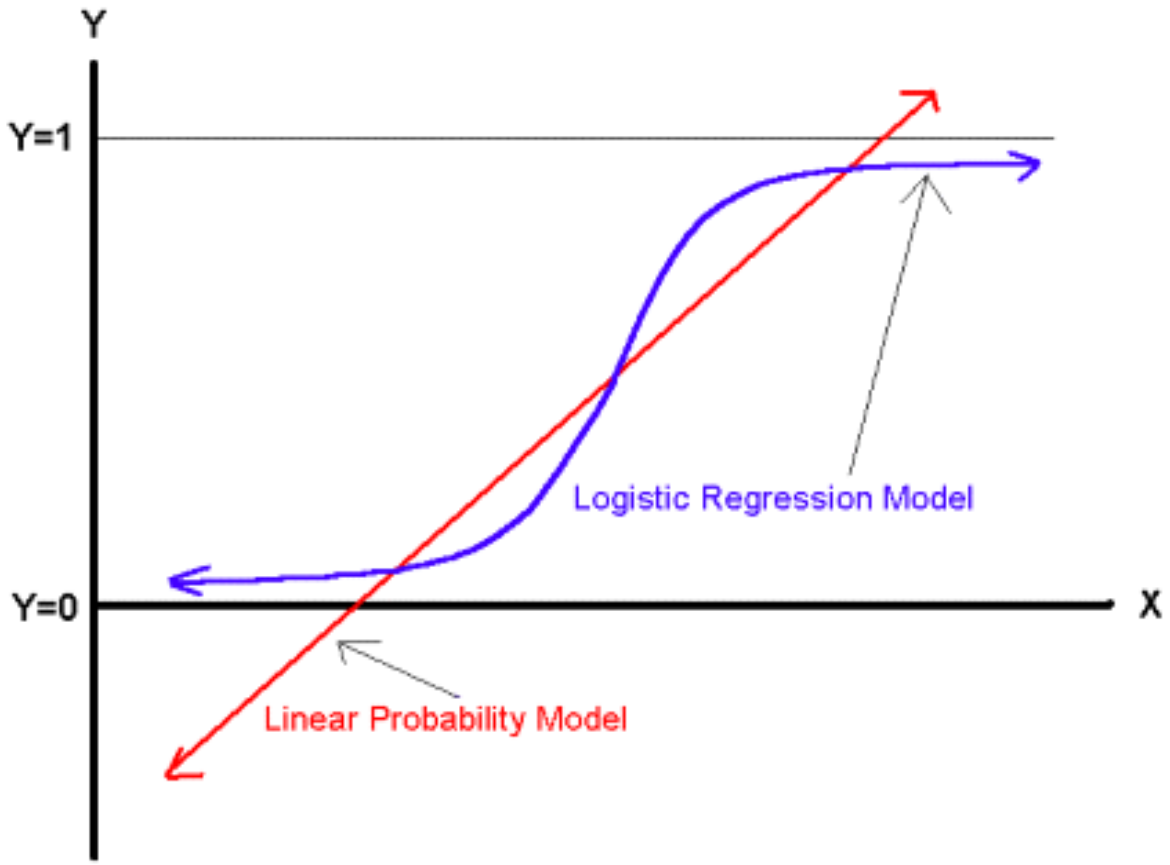
Dasar pemikiran Maximum likelihood (Cont.)

- Jika terdapat 3 bola merah pada kotak, peluang terpilihnya 3 bola merah secara acak adalah

$$\frac{\binom{2}{2}}{\binom{3}{2}} = 1$$

- Oleh karena itu dipilih 3 sebagai penduka dari banyaknya bola merah pada kotak karena 3 merupakan penduga yang memaksimumkan probabilitas dan sampel yang diamati dibandingkan dengan 2 yang probabilitasnya $\frac{1}{3}$ (lebih kecil).
- Kemungkinan terdapat 2 bola merah pada kotak juga benar, tetapi hasil yang diamati memberikan kepercayaan lebih untuk 3 bola merah dalam kotak.
- Contoh ini mengilustrasikan sebuah metode untuk menemukan sebuah penduga yang dapat diaplikasikan pada berbagai situasi. Secara tekis, metode ini disebut Maximum likelihood method.
- Metode ini pertama kali dikenalkan oleh R.A. Fisher pada tahun 1912.
- Metode ini mengasilkan penduga yang sangat baik bagi θ untuk sampel yang besar.

Comparing the LP and Logit Models



Summary

- Two types of regression
  - Linear regression is suitable for continuous data
  - Logistic regression is suitable for categorical data

- But both force us to fit the data with one shape (line or sigmoid) which will often underfit

- Regression is a powerful machine learning technique
  - It provides prediction
  - It offers insight on the relative power of each variable

References

- Hosmer DW, Lemeshow S. Applied logistic regression. Wiley & Sons, New York, 1989
- Richard H. Lathrop, Rachid Salmi, Jean-Claude Desenclos, Thomas Grein, Alain Moren, Introduction to Logistic Regression, 2006, https://www.ics.uci.edu/~rickl/cs-175/
- Luiz Pessoa PY 206 class at Brown University, CS 478 – Tools for Machine Learning and Data Mining, Linear and Logistic Regression, BML BYU Data Mining Lab, http://dml.cs.byu.edu/~
- Toy R. Martinez, Regression, http://axon.cs.byu.edu/~martinez/classes/478/slides/
- Roswita Putri Arcelia Hede, Aris Dwiatmoko, Perbandingan metode kuadrat terkcil dan metode kemungkinanmaksimum dalam pendugaan parameter disribusi weibull dengan dua parameter, Skripsi, Universitas Sanata Dharma, 2016.