Modeling Binary outcome

1. Outcome variable can be binary instead of normally distributed. In biostatistics or epidemiology, we are often interested in the effect of risk factors (x) to a disease (y).

Risk factor value	Disease?
x_1	yes
x_2	no
.	.
.	.
.	.
x_n	no

Table 1. Raw data on risk factor values and disease outcome

2

Table 2. Grouped data on risk factor values and disease outcome.

Risk factor value	Number with disease	Total number	Proportion with disease
x_1	e_1	n_1	r_1
x_2	e_2	n_2	r_2
.	.	.	.
.	.	.	.
.	.	.	.
x_ℓ	e_ℓ	n_ℓ	r_ℓ

2. We are interested in the relationship between risk factors x and r.

Table 3. Prevalent H. pylori and occupational social class amongst men in north Glasgow.

Occupational social class (rank)		Number With *H. pylori*	Number Total	Proportion with *H. pylori*
I	Nonmanual, professional (1)	10	38	0.26
II	Nonmanual, intermediate (2)	40	86	0.46
IIIn	Nonmanual, skilled (3)	36	57	0.63
IIIm	Manual, skilled (4)	226	300	0.75
IV	Manual, partially skilled (5)	83	108	0.77
V	Manual, unskilled (6)	60	73	0.82

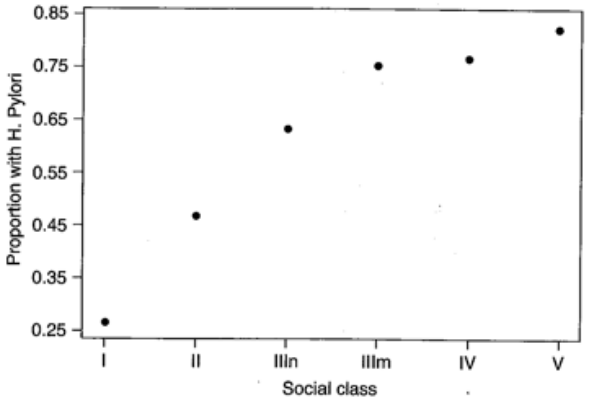


Figure 1.

Table 4. Death by age at baseline; SHHS men.

Age (years)	Number		Percentage dying
	Dying	Total	
40	1	251	0.4
41	12	317	3.8
42	13	309	4.2
43	6	285	2.1
44	10	236	4.2
45	8	254	3.1
46	10	277	3.6
47	12	278	4.3
48	10	285	3.5
49	14	276	5.1
50	15	274	5.5
51	14	296	4.7
52	19	305	6.2
53	36	341	10.6
54	26	305	8.5
55	21	276	7.6
56	28	325	8.6
57	41	302	13.6
58	38	260	14.6
59	49	302	16.2

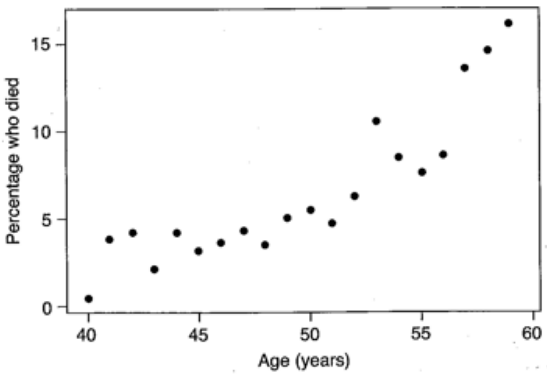


Figure 2.

5

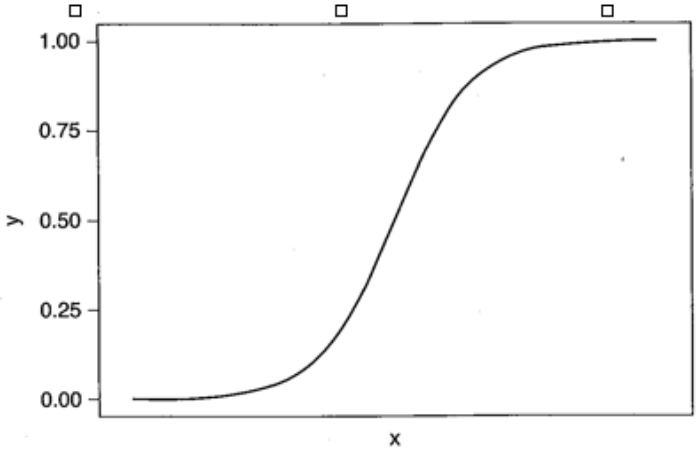
Problems with Linear Regression models

1. The r-x relationship may not be linear
2. Proportions (including risks) must lie between 0 and 1.
3. When observed proportions scan most of this allowable range, the pattern in the scatterplot is generally nonlinear.
4. The tendency toward "squashing up" as proportions approach the asymptotes at 0 or 1.
5. Predicted values of the risk may be outside the valid range:
6. Fitted linear regression model for r regressed on x is given as $r = a + bx$.
7. This can lead to predictions of risks that are negative or are greater than unity, and thus impossible.
8. Fitting a linear regression line to the data in Table 4 gives $r = -25.394 + 0.645 \times age$.
9. If we use this model to predict the risk of death for someone aged 39, the prediction gives $r = -25.394 + 0.645 \times 39 = -0.239$, a negative risk!
10. Similar problems are found with confidence limits for predicted risks within the range of the observed data.
11. The error distribution is not normal. In simple linear regression, we fit the model $r = \alpha + \beta x + \epsilon$, where ϵ arises from a standard normal distribution.
12. r models proportions: proportions are not likely to have a normal distribution; they are likely to be binomial.
13. The inferences drawn from the linear regression would be inaccurate

Logistic regression function

Logistic Regression – Logistic Function

$$y = \left\{1 + \exp(-b_0 - b_1 x)\right\}^{-1}$$



1. The logistic function has an S shape
2. solved the non-linearity problem
3. There is an asymptote at y = 0 and y = 1
4. solved the "out of bound" problem
5. When using logistic function, we assume the data have binomial rather than normal.
6. Solved the assumption of normal error problem
7. The alternative form

$$\log\left(\frac{\hat{r}}{1-\hat{r}}\right) = b_0 + b_1 x$$

8. The left-hand side is called the logit (log of the odds of disease)
9. Logistic regression model postulates a linear relationship between the log odds of disease and the risk factor.
10. The right-hand side is called the linear predictor.

Odds Ratio

If the probability of a success = p, then the **odds in favor of success** = $p/(1-p)$.

Let p_1, p_2 be the underlying probability of success for two groups. The **odds ratio** (***OR***) is defined as

$$OR = \frac{p_1/q_1}{p_2/q_2} = \frac{p_1 q_2}{p_2 q_1} \qquad \text{and is estimated by} \qquad \widehat{OR} = \frac{\hat{p}_1 \hat{q}_2}{\hat{p}_2 \hat{q}_1}$$

Equivalently, if the four cells of the 2×2 contingency table are labeled by a, b, c, d, as they are in Table 13.1, then

$$\widehat{OR} = \frac{[a/(a+b)] \times [d/(c+d)]}{[c/(c+d)] \times [b/(a+b)]} = \frac{ad}{bc}$$

The **disease-odds ratio** is the odds in favor of disease for the exposed group divided by the odds in favor of disease for the unexposed group.

The **exposure-odds ratio** is the odds in favor of being exposed for diseased subjects divided by the odds in favor of being exposed for nondiseased subjects.

Interpretation of logistic regression coefficients

1. Smoking and cardiovascular disease: smoker and disease: 31, smoker and no disease: 1386, nonsmoker and disease: 15, nonsmoker and no disease: 1883.

2.

Parameter	Estimate	Standard error
INTERCEPT	−4.8326	0.2592
SMOKING	1.0324	0.3165

3. $logit = -4.8326 + 1.0324x$, $x = 1$ for smokers and 0 for nonsmokers.

4. The odds ratio for disease, comparing smokers to nonsmokers is $\exp[1.0324(1-0)] = exp[1.0324] = 2.808$

5. Observe that

$$\begin{aligned} \log(\hat{\psi}) &= \log(\hat{odds}_1/\hat{odds}_0) = \log(\hat{odds}_1) - \log(\hat{odds}_0) \\ &= \hat{logit}_1 - \hat{logit}_2 \\ &= b_0 + b_1 x_1 - (b_0 + b_1 x_0) \\ &= b_1(x_1 - x_0) \end{aligned}$$

   Hence $\hat{\psi} = \exp\{b_1(x_1 - x_0)\}$.

6. The estimated standard error of the log odds ratio is 0.3165. An approximate 95% confidence limit for the odds ratio is $\exp[1.0324 \pm 1.96 \times 0.3165] \rightarrow (1.510, 5.221)$

7. Since we know the log odds, we can find odds directly from the fitted logit function.

8. The risk of the disease for smoker is $r = [1 + exp(4.8326 - 1.0324 \times 1]^{-1} = 0.0219 = [1 + exp(-logit)]^{-1}$ implying logit = -3.8002

9. The risk of the disease for nonsmoker is $r = [1 + \exp(4.8326)]^{-1} = 0.0079$

10. The relative risk for smokers to nonsmokers: $0.0219/.0079 = 2.77$