Then the odds ratio for the 15-24 group compared with the non-smoking group is:

$$OR = \frac{445/408}{7/61} = 9.50,$$

so a very large increase in the odds of lung cancer if you smoke 15-25 cigarettes a day. A 95% confidence interval for this odds ratio is created by first calculating the variance of the log odds ratio as:

$$\mathbb{V}[\log(OR)] = \frac{1}{445} + \frac{1}{408} + \frac{1}{7} + \frac{1}{61} = 0.164.$$

Then the 95% CI is given by:

$$\left(\exp(\log(9.50) - 1.96\sqrt{0.164}), \exp(\log(9.50) + 1.96\sqrt{0.164})\right) = (4.29, 21.01).$$

The full set of results for all the smoking levels (relative to non-smokers) are shown in Table 28.

Table 28: Odds ratios and 95% CIs for being diagnosed with lung cancer for different severities of smoking are shown. Smoking severity ranges from smoking 0 to 50+ cigarettes a day.

Cigarettes smoked per day	**Odds ratio**	**95% CI**
0	1.0	
1-4	4.7	(2.0, 11.1)
5-14	7.3	(3.3, 18.1)
15-24	9.5	(4.3, 21.0)
25-49	16.1	(7.2, 36.0)
50+	17.9	(6.9, 46.2)

None of the C.I.s contain 1.0 and there is clear evidence of a 'dose-response' relationship. The evidence of association between smoking and lung cancer in this study is strong. In 1950 there were 4 other case control studies published in the USA, all showing similar results.

4.7 Confounding

Video4.7 - Confounding I

Confounding is a concept apparently so complex, not even Sheldon Cooper from the big bang theory is up to speed with it. We illustrate the idea of confounding with the following example.

Example Suppose you conduct a case-control study into the effect of alcohol consumption on lung cancer, and recruit a set of cases with lung cancer and controls without lung cancer and ask them about their alcohol consumption. Then the estimated odds ratio, say splitting alcohol into drinker / non-drinker would give a very significant result, suggesting that alcohol consumption is a risk factor for lung cancer. However, it is known that no such causal effect exists, so why would the data tell you otherwise? The answer is due to confounding. On average, people who drink are more likely to smoke than people who do not drink, and people who smoke are more likely to get lung cancer. Thus, if you ignore the important confounding variable smoking, then you observe a spurious relationship between alcohol consumption and lung cancer. Thus the golden rule in epidemiology is: **Association does not imply causation.**

Definition **Confounding** refers to the effect of a second variable that either

- wholly or partially accounts for the apparent effect of the exposure.
- masks an underlying true association.

In general, a confounding variable is:

- a risk factor for the disease under study.
- associated with the study exposure (risk factor) and the outcome variable (disease).

Example In a case-control study of oral contraceptive (OC) use and myocardial infarction (MI) the raw odds ratio was estimated as 1.68, so that MI was estimated to be 1.68 times higher among OC users compared with non-users. However, if the data are stratified according to age, as shown in Table 29 below a different pattern appears.

Table 29: Age-specific odds ratios of myocardial infarction risk experienced when taking oral contraceptive vs. not taking oral contraceptive.

Age	25-29	30-34	35-39	40-44	45-49
OR	7.2	8.9	1.5	3.7	3.9

The age-specific odds ratios are well above 1.68 in 4 out of 5 age groups, suggesting that the overall odds ratio estimate of 1.68 is spuriously low. Here age is the confounder. In general a variable which is a confounder can be dealt with:

- at the design stage by matching of cases and controls in a case-control study.
- at the analysis stage by adjustment procedures relying on stratification or regression analysis.

4.7.1 Matching

Definition Matching refers to the pairing of one or more controls to each case on the basis of their similarity with respect to selected variables such as age, sex, marital status, weight, occupation etc. Since cases and matched controls are similar on the matching variables, their differences with respect to disease must be attributed to some other factor.

- Advantages
  - It is easy to understand and justify.
  - It guarantees that cases and controls are comparable on all matching variables.
  - It eliminates the assumption of a particular functional relationship (e.g. linear) made by regression methods.
- Disadvantages
  - There is an increased time and effort to form matches for multiple variables.
  - There is no statistical justification for matching.

4.7.2 Stratification

Video4.8 - Confounding II

Stratification provides a direct method of eliminating biased comparisons that result from confounding. An example of this was shown in a previous section in relation to OC use and MI. The unadjusted odds ratio estimate was 1.68, while the age-specific ratios were 7.2, 8.9, 1.5, 3.7 and 3.9 respectively. Stratification produces one overall summary measure that correctly averages over all the confounder levels. One of the most popular methods to produce such an overall measure is the **Mantel-Haenszel** method. We illustrate this idea for odds ratios via the oral contraceptive example.

Definition Suppose the confounder has been split into G groups or strata, then the 2×2 table for the i^{th} strata is identical to Table 26

The odds ratio for this i^{th} group is

$$OR_i = \frac{a_i d_i}{b_i c_i}.$$

It can be shown that the inverse of the variance of this estimate is

$$w_i = \frac{b_i c_i}{n_i}.$$

Then stratification provides a weighted average of odds ratios, where the weights are given by the inverse of the variances. This is because the bigger the variance (uncertainty) the smaller the weight. Thus the Mantel-Haenszel odds ratio is given by

$$OR_{MH} = \frac{\sum_{i=1}^{G} w_i \frac{a_i d_i}{b_i c_i}}{\sum_{i=1}^{G} w_i}.$$

It is one of the most commonly used methods of adjustment. An estimate of the variance of $\log(OR_{MH})$ is given by

$$\mathbb{V}[\log(OR_{MH})] = \frac{\sum_{i=1}^{G} w_i^2 v_i}{\left(\sum_{i=1}^{G} w_i\right)^2} \quad \text{where} \quad v_i = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}.$$

Thus a 95% Confidence interval can be computed analogously as for unadjusted odds ratios based on a normal approximation, via:

$$\left(\exp\left[\log(OR_{MH}) - 1.96\sqrt{\mathbb{V}[\log(OR_{MH})]}\right] \ , \ \exp\left[\log(OR_{MH}) + 1.96\sqrt{\mathbb{V}[\log(OR_{MH})]}\right]\right).$$

Example Consider again the data from the OC use and MI study. Table 30 below outlines the calculation of the Mantel-Haenszel odds ratio, adjusted for age, and its 95% CI.

Table 30: Several statistics are shown that are needed to compute the age adjusted Mantel-Haenzel odds ratio and 95% CI.

	$\mathbf{25-29}$	$\mathbf{30-34}$	$\mathbf{35-39}$	$\mathbf{40-44}$	$\mathbf{45-49}$
OC	MI / no MI	MI / no MI	MI / no MI	MI / no MI	MI / no MI
Yes	4 / 62	9 / 33	4 / 26	6 / 9	6 / 5
No	2 / 224	12 / 390	33 / 330	65 / 362	93 / 301
$\mathbf{OR_i}$	7.2	8.9	1.5	3.7	3.9
$\mathbf{n_i}$	292	444	393	442	405
$\mathbf{w_i}$	0.425	0.892	2.183	1.324	1.148
$\mathbf{v_i}$	0.771	0.227	0.322	0.296	0.381

Putting these components together yields:

- $OR_{MH} = \dfrac{3.07 + 7.91 + 3.36 + 4.91 + 4.46}{0.43 + 0.89 + 2.18 + 1.32 + 1.15} = 3.97.$
- $\mathbb{V}[\log(OR_{MH})] = \dfrac{\sum_{i=1}^{G} w_i^2 v_i}{\left(\sum_{i=1}^{G} w_i\right)^2} = \dfrac{2.876}{(5.972)^2} = 0.0806.$
- 95% CI: (2.28,6.93).

These results compare to the raw unadjusted odds ratio of 1.68 (1.10, 2.57).

Note One can of course extend this approach to more than one confounder, but the number of separate odds ratios that need to be computed quickly gets out of control.

4.8 Regression modelling

Video4.9 - Regression modelling I

The most common approach for quantifying the effect of an exposure on disease risk whilst allowing for confounders is regression modelling. Regression modelling is a vast topic covered in many of your courses (e.g. Linear Models, Generalised Linear Models, Regression Modelling M, etc), and we only give a brief flavour here. Consider a study with $i = 1, \ldots, n$ individuals or study units (e.g. small areas for an ecological study), then a regression analysis requires three basic components:

- A response variable Y_i, which in this epidemiological context is a measure of disease. This variable could be:
  - Continuous - such as blood pressure, weight.
  - Binary - such as disease presence or absence (you either have a disease or not).
  - Count - such as the number of people in each small-area with the disease.
- A set of p covariates for each individual / study unit i, $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})$, where the '1' corresponds to an intercept term in the regression model. These risk factors include the exposure of interest and confounding factors, and can contain both continuous and factor variables.
- The covariates each have an associated regression parameter $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$, where β_j is the effect of the j^{th} covariate on disease.

4.8.1 A model for continuous data

The most well known model for continuous response variables is the normal linear model, which is given by:

$$Y_i \sim \mathsf{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2) \qquad i = 1, \ldots, n,$$

where σ^2 is the observation error variance. Estimation of $(\boldsymbol{\beta}, \sigma^2)$ is achieved using maximum likelihood, see the Linear Models / Regression Modelling M courses for more information. In terms of interpretation, β_j represents the effect of the j^{th} covariate on the response, so that:

- If the j^{th} covariate is continuous then a 1 unit increase in x_j is expected to result in a β_j unit increase in the response variable Y_j.
- If the j^{th} covariate is a level in a factor, then its coefficient β_j represents the expected increase in the response variable Y_j from moving from the baseline level of the factor to the level represented by the j^{th} covariate.

4.8.2 A model for binary data

Binary regression models are typically used to model individual-level data, where the response for each individual is whether or not they have the disease under study. Let Y_i be the binary indicator variable for $i = 1, \ldots, n$ individuals, where $Y_i = 1$ if individual i has the disease and $Y_i = 0$ if they do not. Then the normal linear model is not appropriate, as the response variable is in the set $\{0, 1\}$ rather than being continuous. Normal linear models can be extended naturally to account for binary response variables. Consider the alternative formulation of the normal linear model:

$$Y_i \sim \mathsf{N}(\mu_i, \sigma^2) \qquad i = 1, \ldots, n,$$

$$\mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

Then a commonly used model for binary data are called **logistic regression**, and follows this general form:

$$Y_i \sim \text{Bernoulli}(\theta_i) \qquad i = 1, \ldots, n,$$

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

Here θ_i is the probability of disease for individual i, that is $\mathbb{P}(Y_i = 1) = \theta_i$, and is also the mean $\mathbb{E}(Y_i) = \theta_i$.

Notes

- The normal likelihood model has been replaced by the Bernoulli likelihood model as the data are binary.
- Additionally, $\mathbb{E}(Y_i) \neq \mathbf{x}_i^\top \boldsymbol{\beta}$ as in the linear model, and instead the **logistic** transformation $f(x) = \log\left(\frac{x}{1-x}\right)$ relates the expected value of the response to the linear predictor $\mathbf{x}_i^\top \boldsymbol{\beta}$. This is because $\theta_i \in [0, 1]$ where as $\mathbf{x}_i^\top \boldsymbol{\beta}$ can take values in the whole real line, hence the transformation $\log\left(\frac{\theta_i}{1-\theta_i}\right)$ which can also take values in the whole real line.

- The probability of disease can be obtained for individual i as

$$
\begin{aligned}
\log\left(\frac{\theta_i}{1-\theta_i}\right) &= \mathbf{x}_i^\top \boldsymbol{\beta} \\
\frac{\theta_i}{1-\theta_i} &= \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \\
\theta_i &= \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) - \theta_i \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \\
\theta_i(1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta})) &= \exp(\mathbf{x}_i^\top \boldsymbol{\beta}) \\
\theta_i &= \frac{\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^\top \boldsymbol{\beta})}.
\end{aligned}
$$

- The parameter vector $\boldsymbol{\beta}$ can be estimated using maximum likelihood, and 95% confidence intervals constructed based on a normal approximation.
- In terms of interpretation, **odds ratios** are commonly used in logistic regression models. Recall from Measuring the association between a risk factor and disease that the odds of disease for individual i are the likelihood of disease divided by likelihood of no disease, which of individual i is given by

$$
\frac{\theta_i}{1-\theta_i} \quad = \quad \exp(\mathbf{x}_i^\top \boldsymbol{\beta}).
$$

Then the odds ratio for individual i of a single covariate x_p increasing by M is:

$$
\begin{aligned}
OR &= \frac{\text{odds}(x_{ip}+M)}{\text{odds}(x_{ip})} \\
&= \frac{\exp(\beta_1 + x_{i2}\beta_2 + \ldots + (x_{ip}+M)\beta_p)}{\exp(\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p)} \\
&= \exp(M\beta_p).
\end{aligned}
$$

So the odds ratio is straightforward to calculate and does not depend on the individual i. A similar result holds for categorical covariates.

Example A study was set up to identify the risk factors for the development of Coronary Heart Disease (CHD) in a cohort of 742 men aged 40-49 years and free of disease at the initial examination. The cohort was followed up for 12 years, during which time 88 men (11.9%) developed symptoms of CHD. Table 31 summarises the results of a logistic regression analysis.

Table 31: The results of a logistic regression model as one would expect to see from standard R output. The response is CHD (yes/no).

Variable	Parameter	Parameter estimate	Standard error	p-value
intercept	β_0	-13.2573		
x_1, Age	β_1	0.1216	0.0437	0.006
x_2, Cholestoral	β_2	0.0070	0.0025	0.005
x_3, Systolic BP	β_3	0.0068	0.006	0.26
x_4, Weight	β_4	0.0257	0.0091	0.005
x_5, Haemoglobin	β_5	-0.0010	0.0098	0.9
x_6, Smoking habit	β_6	0.4223	0.1031	<0.0001
x_7, ECG abnormality	β_7	0.8206	0.4009	0.04

A number of key quantities can be obtained from this logistic regression analysis.

- The probability of getting CHD within 12 years for any combination of covariates can be obtained as follows. Consider a man with $\mathbf{x} = (45, 210, 130, 100, 120, 0, 0)$. Based on this model, his 12-year risk of developing CHD is:

$$\theta_x = \frac{\exp(-13.2573 + 0.1216 \times 45 + \ldots + 0.8206 \times 0)}{1 + \exp(-13.2573 + 0.1216 \times 45 + \ldots + 0.8206 \times 0)} = 0.0483.$$

- Odds ratios as previously described are easy to compute. So for example, the odds ratio for getting CHD with ECG abnormality compared to without ECG abnormality is:

$$OR = \exp(0.8206) = 2.27.$$

So one has a 2.27 higher odds of getting CHD if they have an ECG abnormality.

4.8.3 A model for count data

Video4.10 - Regression modelling II

Count data regression models are typically used to model population-level data, where the response in each unit is the number of cases of disease from the population who live there. Consider a study region split into n areas, then the number of cases of disease from area i are denoted by Y_i. As the response data are non-negative counts, a Poisson likelihood model is used in place of a normal likelihood, yielding the model:

$$Y_i \sim \text{Poisson}(\lambda_i) \qquad i = 1, \ldots, n,$$

$$\log(\lambda_i) = \mathbf{x}_i^\top \boldsymbol{\beta}^*.$$

However, the number of disease cases observed in each area will depend on each areas population size and its age and sex structure. Therefore indirect standardisation is used to compute the expected number of disease cases, denoted by E_i for area i. The model above is modified to give

$$Y_i \sim \text{Poisson}(E_i R_i) \qquad i = 1, \ldots, n,$$

$$\log(R_i) = \log(\frac{\lambda_i}{E_i}) = \mathbf{x}_i^\top \boldsymbol{\beta}.$$

Notes

- Here R_i is the risk of disease in the i^{th} study unit, and has been standardised indirectly and therefore represents the SMR. The log-linear relationship between R_i and the linear predictor $\mathbf{x}_i^\top \boldsymbol{\beta}$ is used as a Poisson mean must be positive where as the linear predictor could take on any value.
- The parameter vector $\boldsymbol{\beta}$ can be estimated using maximum likelihood, and 95% confidence intervals constructed based on a normal approximation.
- In terms of interpretation *relative risks* are commonly used in Poisson log-linear models, which is the ratio of risks of disease with differing exposure levels. For example, suppose variable x_p increases to $x_p + M$, then the relative risk for individual or study unit i would be:

$$RR = \frac{\text{Risk of disease based on } x_{ip} + M}{\text{Risk of disease based on } x_{ip}}$$

$$= \frac{R_i(x_{ip} + M)}{R_i(x_{ip})}$$

$$= \frac{\exp(\beta_0 + x_{i1}\beta_1 + \ldots + (x_{ip} + M)\beta_p)}{\exp(\beta_0 + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p)}$$

$$= \exp(M\beta_p).$$

  Again this does not depend on the i^{th} individual or study unit and is the same for all individuals / units. A similar result holds for categorical covariates.

Example The data come from a study investigating the spatial pattern in cancer risk in Greater Glasgow, Scotland, between 2001 and 2005. The data for this study are publicly available, and come from the Scottish Statistics database. The study region is the GGCHB, which contains the largest city in Scotland (Glasgow) as well as the surrounding area. A small number of covariates are available to describe the spatial variation in cancer risk across Greater Glasgow.

- A modelled estimate of the percentage of the population in each IG who smoke.
- The percentage of school children from ethnic minorities (i.e. non-white), which is used here as a proxy measure of the ethnic make-up of each area.
- The estimated annual mean concentration of particulate matter air pollution in 2001, which is measured as PM$_{10}$ (small solid and liquid particles less than 10 microns in size).

The results of fitting a Poisson model to these data are shown below, where the expected number of cases was included in the model as an offset on the log scale.

```
Call:
glm(formula = Y_all ~ offset(log(E_all)) + pm10 + smoke + ethnic,
family = poisson, data = data2)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.5470147  0.0565970  -9.665  < 2e-16 ***
pm10         0.0227687  0.0035544   6.406  1.5e-10 ***
smoke        0.0082125  0.0006203  13.239  < 2e-16 ***
ethnic      -0.0049302  0.0005853  -8.423  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Using the relative risk transformation described above the relative risk for a one unit increase in each covariate is as follows:

- PM10 - 1.023 (1.016, 1.030).
- Smoke - 1.008 (1.007, 1.009).
- Ethnic - 0.995 (0.994, 0.996).

So, why does smoking have a smaller relative risk than PM10? Well, the standard deviation in smoking percentage across GGCHB is 9.63, where as for PM10 it is 1.79, therefore a 1 unit increase in smoking is small where as comparatively for PM10 it is much larger. Note, those standard deviations can be obtained from open data; they are not the standard deviations from the output above. The standard deviations from the model output relate to the model parameters. If we compute relative risks for a 1 standard deviation increase we have:

- PM10 - 1.042 (1.029, 1.055).
- Smoke - 1.082 (1.070 1.095).
- Ethnic - 0.941 (0.928, 0.955).

Interpretability of scale is crucial here.