

## Lab 2: Solutions

### 1. The Diet Dataset

**File:** diet

**Description:** This dataset, derived from a prospective cohort pilot study, comprises dietary and health information from 337 adult male subjects, collected over a two-week period. The primary outcome of interest is Coronary Heart Disease (CHD), represented by the binary variable 'chd', where 1 indicates the presence of CHD and 0 its absence.

#### 1.1. Data Exploration

1. Read the data into Stata and examine the variables.

```
. use diet, clear
. desc
```

```
obs:      337      Diet data with dates
vars:      13      7 May 2008 15:40
size:    15,839
```

---

variable name	storage type	display format	value label	variable label
id	float	%9.0g		Subject identity number
doe	long	%dDmCY		Date of entry
dox	long	%dDmCY		Date of exit
chd	float	%9.0g		Outcome: 1= chd, 0 otherwise
dob	long	%dDmCY		Date of birth
job	int	%8.0g		Occupation
month	byte	%8.0g		month of survey
energy	float	%9.0g		Total energy (1000kcal/day)
height	float	%9.0g		Height (cm)
weight	float	%9.0g		Weight (kg)
fat	float	%9.0g		Total fat (g/day)
fibre	float	%9.0g		Total fibre (g/day)
hieng	float	%9.0g		Indicator for energy > 2.75

---

The dataset contains **337 observations** and **13 variables**. Key variables include subject ID, dates of entry and exit, CHD outcome, date of birth, occupation, and various dietary and physical measurements.

2. Determine the number of CHD cases in the dataset. Record this number for future reference.

```
. tab chd
```

Outcome: 1= chd, 0 otherwise	Freq.	Percent	Cum.
0	291	86.35	86.35
1	46	13.65	100.00
Total	337	100.00	

There are 46 CHD cases out of 337 subjects, representing 13.6% of the sample.

### 3. Examine the date variables

```
. list id doe dox chd in 1/10
```

	id	doe	dox	chd
1.	1	16Aug1964	01Dec1976	0
2.	2	16Dec1964	01Dec1976	0
3.	3	16Nov1965	01Dec1976	0
4.	4	16Sep1965	01Dec1976	0
5.	5	16Sep1965	31Mar1976	0
6.	6	16Mar1965	31Aug1968	0
7.	7	16Nov1958	01Dec1976	0
8.	8	16May1965	01Dec1976	0
9.	9	16Feb1959	10Jan1962	0
10.	10	16Jul1964	16May1974	0

The date variables (doe for date of entry and dox for date of exit) are stored in STATA format as days since January 1, 1960. These are treated as numbers for calculations but displayed in standard date format for output.

## 1.2. Setting up Survival-Time Data

### 1. Set the survival-time (st) variables

```
. stset dox, failure(chd) origin(doe) scale(365.25)
```

```

failure event:  chd != 0 & chd < .
obs. time interval:  (origin, dox]
exit on or before:  failure
t for analysis:  (time-origin)/365.25
origin:  time doe
```

```

337 total observations
0 exclusions
```

```

337 observations remaining, representing
46 failures in single-record/single-failure data
4603.669 total analysis time at risk and under observation
              at risk from t =              0
earliest observed entry t =              0
last observed exit t = 20.04107
```

### Interpretation of stset output

The dataset contains 337 subjects, among whom 46 CHD events were observed. The total person-time at risk is 4,603.67 years, and the maximum follow-up time is approximately 20 years. No observations were excluded by the stset declaration.

### Role of stset

The stset command declares the data to be survival-time data and specifies how analysis time (`_t` and `_t0`) and the failure indicator (`_d`) are to be constructed. Stata then checks for internal consistency of the declaration. Running `streset` recomputes these quantities and reruns the same checks.

**a. Failure event:** The option `failure(chd)` defines CHD as the failure event. Any non-zero, non-missing value of `chd` is treated as a failure, and 0 or missing as censoring. More complex definitions (e.g., multiple codes) can be specified if needed.

**b. Entry and exit times:** This dataset has one record per subject, so entry and exit times correspond to the beginning and end of each subject's follow-up period. Because `enter()` is not specified, subjects are assumed to come under observation at the time implied by the origin.

Exit time is given directly by the time variable declared in `stset`, here `dox`. Subjects without a failure are censored at their last observed time.

**c. Time scale:** `origin(doe)` sets the origin of analysis time to each subject's date of entry into the study. `scale(365.25)` rescales the time scale from days to years. This does not change the amount of person-time but changes its units.

**d. Analysis time:** Given the specification:  $t = \frac{dox - doe}{365.25}$ , analysis time is measured in years since study entry, starting at `_t0 = 0` and ending at `_t`.

### Practical Implications of this Setup:

- Survival times and incidence rates can be interpreted per year.
- Right-censored observations (subjects who do not fail before their last follow-up time) are handled naturally through the failure indicator.
- Differences in entry times and follow-up lengths are properly represented because each subject contributes observation time from `_t0` to `_t`.
- This `stset` configuration provides the structure needed for subsequent survival analyses (e.g., rate estimation, Kaplan-Meier curves, Cox regression).

## 2. Examine the newly created variables

```
. list id _t0 _t _d _st in 1/10
```

	id	_t0	_t	_d	_st
1.	1	0	12.29295	0	1
2.	2	0	11.958932	0	1
3.	3	0	11.041752	0	1
4.	4	0	11.208761	0	1
5.	5	0	10.537988	0	1
6.	6	0	3.4606434	0	1
7.	7	0	18.042437	0	1
8.	8	0	11.545517	0	1
9.	9	0	2.899384	0	1
10.	10	0	9.8316222	0	1

**\_t0:** Time of entry on the analysis time scale (in years since origin)

**\_t:** Time of exit on the analysis time scale (in years since origin)

**\_d:** Failure indicator (1 = event occurred, 0 = censored)

**\_st:** Indicator for whether the record is included in the analysis (1 = included, 0 = excluded)

The difference between **\_t** and **\_t0** represents the duration of follow-up for each subject. In this study, all **\_t0** values are 0 because we set the origin to the date of entry (doe), but this isn't always the case in all survival analyses.

## 3. Change the time scale from time-since-entry to age.

```
. stset dox, failure(chd) origin(dob) enter(doe) scale(365.25)
```

```
failure event:  chd != 0 & chd < .
obs. time interval:  (origin, dox]
enter on or after:  time doe
exit on or before:  failure
t for analysis:  (time-origin)/365.25
origin:  time dob
```

---

```
337 total observations
0 exclusions
```

---

```
337 observations remaining, representing
46 failures in single-record/single-failure data
4603.669 total analysis time at risk and under observation
at risk from t = 0
earliest observed entry t = 30.07529
last observed exit t = 69.99863
```

The time scale was then changed from time-since-entry to age by setting the origin to the date of birth (dob) and specifying entry at the date of entry into the study (doe). In this specification,

analysis time is measured in years of age: `_t0` is the age at study entry, and `_t` is the age at exit (event or censoring). Although the time scale has changed, the person-time at risk for each subject remains `_t - _t0`, so the total person-time in the study is identical to the previous setup; only the interpretation of the time axis (age instead of time since entry) has changed.

```
. list id _t0 _t _d _st in 1/20
```

	id	_t0	_t	_d	_st
1.	1	49.615332	61.908282	0	1
2.	2	50.537988	62.49692	0	1
3.	3	58.784394	69.826146	0	1
4.	4	58.726899	69.935661	0	1
5.	5	59.460643	69.998631	0	1
6.	6	50.98152	54.442163	0	1
7.	7	45.138946	63.181383	0	1
8.	8	50.428474	61.97399	0	1
9.	9	67.099247	69.998631	0	1
10.	10	60.167009	69.998631	0	1

### 1.3 Preliminary Analysis

1. Examine the distribution of the high energy intake (`hieng`) variable.

```
. tab hieng
```

Indicator for energy > 2.75	Freq.	Percent	Cum.
0	155	45.99	45.99
1	182	54.01	100.00
Total	337	100.00	

The variable `hieng` is a binary exposure indicator, equal to 1 for total energy intake >2.75 Mcal/day and 0 otherwise. In this dataset, 54% of participants had `hieng` = 1, providing adequate exposure contrast.

2. Calculate CHD rates using the `strate` command.

```
. strate, per(1000)
```

```

failure _d:   chd
analysis time _t: (dox-origin)/365.25
              origin: time dob
enter on or after: time doe

```

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals  
(337 records included in the analysis)

D	Y	Rate	Lower	Upper
46	4.6037	9.9920	7.4843	13.3400

The strate command, with the per(1000) option, calculates the overall incidence rate of CHD per 1000 person-years.

### Key results:

Number of CHD cases (D): 46

Total person-time at risk (Y): 4,603.7 person-years (4.6037 \* 1000)

Incidence rate: 9.9920 per 1000 person-years (95% CI: 7.48–13.34).

**Interpretation:** On average, we expect about 10 (95% CI: 7.48–13.34) new cases of CHD for every 1000 person-years of follow-up in this population.

The strate command can also stratify rates by levels of categorical variables when specified in the command's varlist. For example,

```
. strate hieng, per(1000)
```

```

failure _d:   chd
analysis time _t: (dox-origin)/365.25
              origin: time dob
enter on or after: time doe

```

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals  
(337 records included in the analysis)

hieng	D	Y	Rate	Lower	Upper
0	28	2.0594	13.5960	9.3875	19.6912
1	18	2.5442	7.0748	4.4574	11.2291

3. Create a new categorical variable called htgrp for height groups using the cut points (150, 170, 175, 180, 195 cm).

```
. egen htgrp = cut(height), at(150, 170, 175, 180, 195) icodes
```

```
. label var htgrp "Height category"
. label def hg      0 "[150-170cm]" 1 "[170-175cm]" ///
                  2 "[175-180cm]" 3 "[180-195cm]"
. label values htgrp hg
```

```
. tab htgrp
```

Height category	Freq.	Percent	Cum.
[150-170cm)	92	27.71	27.71
[170-175cm)	102	30.72	58.43
[175-180cm)	83	25.00	83.43
[180-195cm)	55	16.57	100.00
Total	332	100.00	

A new categorical variable 'htgrp' was created for height groups using the cut points (150, 170, 175, 180, 195 cm). This allows for analysis of CHD rates across different height categories.

```
. bysort htgrp: sum height, det // check if the recording was successful
```

```
-> htgrp = [150-170cm)
```

Height (cm)				
Percentiles	Smallest			
1% 152.4	152.4			
5% 157.988	153.67			
10% 160.02	157.48	Obs		92
25% 163.83	157.734	Sum of Wgt.		92
50% 166.37		Mean		165.5602
	Largest	Std. Dev.		3.530403
75% 167.7924	169.3926			
90% 168.91	169.545	Variance		12.46375
95% 169.1894	169.799	Skewness		-1.392987
99% 169.926	169.926	Kurtosis		5.127077

```
-> htgrp = [170-175cm)
```

Height (cm)				
Percentiles	Smallest			
1% 170.0022	170.0022			
5% 170.18	170.0022			
10% 170.18	170.18	Obs		102
25% 171.45	170.18	Sum of Wgt.		102
50% 172.72		Mean		172.277
	Largest	Std. Dev.		1.338051
75% 173.482	173.99			
90% 173.99	174.1932	Variance		1.790381
95% 173.99	174.498	Skewness		-.1539058
99% 174.498	174.625	Kurtosis		1.892441

```
-> htgrp = [175-180cm)
```

Height (cm)				
	Percentiles	Smallest		
1%	175.006	175.006		
5%	175.26	175.006		
10%	175.26	175.26	Obs	83
25%	175.895	175.26	Sum of Wgt.	83
50%	176.53		Mean	177.0533
		Largest	Std. Dev.	1.355743
75%	177.8	179.07		
90%	179.07	179.07	Variance	1.838039
95%	179.07	179.07	Skewness	.1356182
99%	179.07	179.07	Kurtosis	1.75807

```
-> htgrp = [180-195cm)
```

Height (cm)				
	Percentiles	Smallest		
1%	180.34	180.34		
5%	180.34	180.34		
10%	180.34	180.34	Obs	55
25%	180.34	180.34	Sum of Wgt.	55
50%	182.88		Mean	182.8292
		Largest	Std. Dev.	2.570328
75%	184.15	187.96		
90%	186.69	187.96	Variance	6.606587
95%	187.96	189.23	Skewness	1.049664
99%	190.5	190.5	Kurtosis	3.490751

4. Calculate and compare CHD rates across the different height groups (htgrp) using the `strate` command.

```
. strate htgrp, per(1000)
```

```
failure _d: chd
analysis time _t: (dox-origin)/365.25
origin: time dob
enter on or after: time doe
```

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals  
(332 records included in the analysis)

htgrp	D	Y	Rate	Lower	Upper
[150-170cm)	19	1.1532	16.4760	10.5093	25.8304
[170-175cm)	15	1.3212	11.3537	6.8447	18.8328
[175-180cm)	11	1.1496	9.5689	5.2992	17.2785
[180-195cm)	0	0.9097	0.0000	.	.

CHD rates were calculated across the different height groups. The crude rates appeared lower in the taller categories, but these differences are based on small numbers and should be regarded as



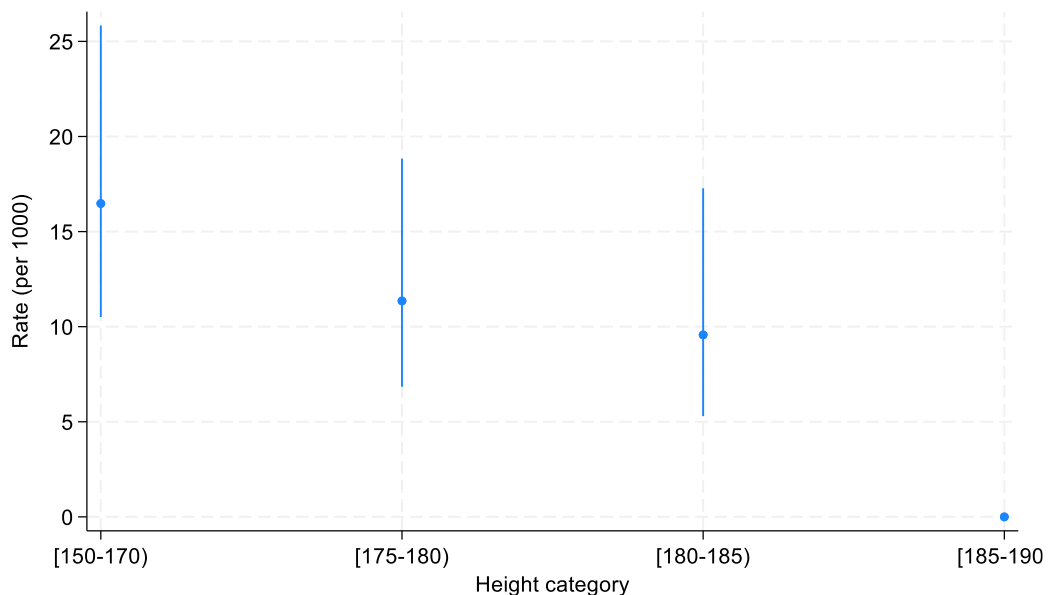
descriptive and exploratory. Specifically:

- 150–169 cm: 16.48 per 1000 person-years
- 170–174 cm: 11.35 per 1000 person-years
- 175–179 cm: 9.57 per 1000 person-years
- $\geq 180$  cm: 0 per 1000 person-years (no cases observed in this group)

Across height groups, the crude CHD rates appear lower among taller men, but this pattern is based on sparse data and no formal trend test has been performed. In particular, the zero rate in the tallest group ( $\geq 180$  cm) is compatible with random variation due to the small number of subjects rather than complete protection. Moreover, height is likely correlated with other factors (for example, socioeconomic and health status) that are not controlled for here, so these results should be interpreted as exploratory rather than evidence of a dose–response relationship.

5. Create an appropriate graph to visualize how CHD rates change across height groups.

```
. strate htgrp, per(1000) graph
```



This graph displays the crude CHD rates by height group; the apparent decrease with increasing height, including the zero rate in the tallest group, should be interpreted cautiously because of sparse data and lack of adjustment for confounding.

#### 1.4 Rate Ratios

1. Calculate rate ratios for high energy intake

```
. stmh hieng
```

```
failure _d: chd
analysis time _t: (dox-origin)/365.25
origin: time dob
enter on or after: time doe
```

Maximum likelihood estimate of the rate ratio  
comparing hieng==1 vs. hieng==0

RR estimate, and lower and upper 95% confidence limits

RR	chi2	P>chi2	[95% Conf. Interval]	
0.520	4.84	0.0277	0.288	0.941

### Interpretation:

The estimated rate ratio for CHD comparing men with high energy intake (>2.75 Mcal/day) to those with lower intake (≤2.75 Mcal/day) is 0.52 (95% CI 0.29–0.94;  $p = 0.028$ ). This estimate is compatible with a lower CHD incidence rate in the high-energy group. However, this is an observational, unadjusted comparison, and the apparent inverse association may be partly or wholly explained by confounding factors (for example differences in physical activity, overall dietary pattern, occupational demands, or underlying health status) that are not controlled for in this analysis.

The observed inverse association should therefore be interpreted cautiously. The estimated rate ratio may reflect confounding rather than a causal effect of energy intake. Several factors could plausibly influence both energy intake and CHD incidence, including:

- physical activity, which increases energy requirements and is typically inversely associated with CHD;
- overall dietary pattern, since total energy intake does not capture nutrient composition or diet quality;
- occupational demands, especially physically demanding work that requires higher caloric intake;
- underlying health status, which may influence both eating patterns and CHD risk.

**stmh hieng:** By default, it compares group 1 (high energy) vs group 0 (low energy)

$$RR = \frac{\text{Rate}(\text{hieng} = 1)}{\text{Rate}(\text{hieng} = 0)}$$

```
. stmh hieng, c(0, 1)
```

```
failure _d: chd
analysis time _t: (dox-origin)/365.25
origin: time dob
enter on or after: time doe
```

Maximum likelihood estimate of the rate ratio  
comparing hieng==0 vs. hieng==1

RR estimate, and lower and upper 95% confidence limits

RR	chi2	P>chi2	[95% Conf. Interval]	
1.922	4.84	0.0277	1.063	3.474

**stmh hieng, c(0,1):** The c(0,1) option reverses the comparison. That is, it compares group 0 (low energy) vs group 1 (high energy).

$$RR = \frac{\text{Rate}(\text{hieng} = 0)}{\text{Rate}(\text{hieng} = 1)}$$

#### Interpretation:

The Mantel–Haenszel rate ratio of 1.922 indicates that the incidence rate of CHD among men with *low energy intake* ( $\leq 2.75$  Mcal/day) is approximately 1.9 times higher than that among men with *high energy intake* ( $> 2.75$  Mcal/day).

## 2. Create a new variable with multiple energy intake levels

```
. egen eng3=cut(energy), at(1.5, 2.5, 3.0, 4.5) icodes
. tab eng3
```

eng3	Freq.	Percent	Cum.
0	75	22.26	22.26
1	150	44.51	66.77
2	112	33.23	100.00
Total	337	100.00	

**Note:** The icodes option tells Stata to assign **integer codes (0, 1, 2, ...)** to each cut interval instead of using the actual cutpoint values.

This is especially useful for trend analysis and simplifies interpretation when modeling ordered exposure levels.

## 3. Analyze rates for different energy levels

```
. strate eng3, per(1000)
```

```
failure _d: chd
analysis time _t: (dox-origin)/365.25
origin: time dob
enter on or after: time doe
```

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals  
(337 records included in the analysis)

eng3	D	Y	Rate	Lower	Upper
0	16	0.9466	16.9020	10.3547	27.5892
1	22	2.0173	10.9059	7.1810	16.5629
2	8	1.6398	4.8787	2.4398	9.7555

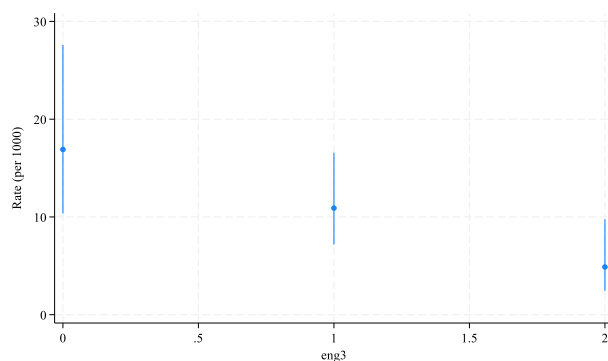
The analysis using eng3 (three categories of energy intake) offers a clearer view of the relationship between energy intake and CHD risk. By examining three distinct levels (1.5-2.5, 2.5-3.0, and 3.0-4.5 Mcal/day), we can assess whether the relationship is linear or shows threshold effects. The crude CHD rates decrease across the three energy-intake categories. This pattern is consistent with a monotonic trend but is based on unadjusted data and should be interpreted as descriptive rather than as evidence of a causal dose–response relationship.

**. strate eng3, per(1000) graph**

```
failure _d: chd
analysis time _t: (dox-origin)/365.25
origin: time doe
```

Estimated rates (per 1000) and lower/upper bounds of 95% confidence intervals  
(337 records included in the analysis)

eng3	D	Y	Rate	Lower	Upper
0	16	0.9466	16.9020	10.3547	27.5892
1	22	2.0173	10.9059	7.1810	16.5629
2	8	1.6398	4.8787	2.4398	9.7555



Note that with the option graph we can plot the actual rates. Therefore, this graph offers a graphical inspection of the actual difference in the rates.

#### 4. Compare CHD rates between different levels of energy intake

```
. stmh eng3, c(1,0)
```

```
failure _d: chd
analysis time _t: (dox-origin)/365.25
origin: time dob
enter on or after: time doe
```

Maximum likelihood estimate of the rate ratio  
comparing eng3==1 vs. eng3==0

RR estimate, and lower and upper 95% confidence limits

RR	chi2	P>chi2	[95% Conf. Interval]	
0.645	1.81	0.1789	0.339	1.229

```
. stmh eng3, c(2,0)
```

```
failure _d: chd
analysis time _t: (dox-origin)/365.25
origin: time dob
enter on or after: time doe
```

Maximum likelihood estimate of the rate ratio  
comparing eng3==2 vs. eng3==0

RR estimate, and lower and upper 95% confidence limits

RR	chi2	P>chi2	[95% Conf. Interval]	
0.289	9.35	0.0022	0.124	0.674

#### 5. Investigate the effect of high energy intake on CHD rate

```
. stmh eng3
```

```
failure _d: chd
analysis time _t: (dox-origin)/365.25
origin: time dob
enter on or after: time doe
```

Score test for trend of rates with eng3  
with an approximate estimate of the  
rate ratio for a one unit increase in eng3

RR estimate, and lower and upper 95% confidence limits

RR	chi2	P>chi2	[95% Conf. Interval]	
0.548	8.98	0.0027	0.370	0.812

With `stmh eng3`, `eng3` is treated as an ordered exposure. The command performs a score test for linear trend in the log CHD rate across the three `eng3` categories and reports the corresponding

trend rate ratio. The estimate  $RR = 0.548$  ( $\chi^2 = 8.98$ ,  $p = 0.0027$ ) means that, **under a log-linear trend assumption**, moving up one eng3 category multiplies the CHD incidence rate by about 0.55 on average, indicating a decreasing rate with higher energy-intake levels.

```
. stmh energy
```

```
failure _d: chd
analysis time _t: (dox-origin)/365.25
origin: time dob
enter on or after: time doe
```

```
Score test for trend of rates with energy
with an approximate estimate of the
rate ratio for a one unit increase in energy
```

```
RR estimate, and lower and upper 95% confidence limits
```

RR	chi2	P>chi2	[95% Conf. Interval]	
0.351	9.89	0.0017	0.183	0.674

Here we assess the effect of 1 unit (1 Mcal) increase in the actual value of energy intake. The rate ratio of 0.351 indicates a reduction with 1 Mcal increase in the energy intake. As expected, this is statistically significant also.

Interpretation of the **stmh** command in Stata

Type of exposure variable	What stmh calculates	Null hypothesis tested ( $\chi^2$ )	Interpretation of the rate ratio
<b>Binary</b> (e.g. hieng)	Mantel-Haenszel rate ratio (MH RR)	$H_0$ : $RR = 1$ vs $H_1$ : $RR \neq 1$	Compares the CHD rate between exposed and unexposed groups (adjusted for strata if specified)
<b>Ordered categorical or continuous</b> (e.g. eng3, energy)	Score test for trend in rates	$H_0$ : no monotonic trend in rates across exposure levels vs $H_1$ : a monotonic trend exists	The reported RR <b>approximates</b> the proportional change in rate per one-unit increase in the exposure (according to the variable's coding or assigned scores).

### 1.5 Controlling for Confounding

1. Examine the effect of high energy intake on CHD rate, controlling for job.

```
. stmh hieng job
```

```

failure _d:  chd
analysis time _t:  (dox-origin)/365.25
              origin:  time dob
enter on or after:  time doe

```

Mantel-Haenszel estimate of the rate ratio  
 comparing hieng==1 vs. hieng==0  
 controlling for job

RR estimate, and lower and upper 95% confidence limits

RR	chi2	P>chi2	[95% Conf. Interval]	
0.525	4.71	0.0299	0.291	0.949

Here the aim is to explore the effect of energy intake on the rates of CHD, adjusting for any effect from the variable job.

### Interpretation:

Adjustment for occupation had minimal effect compared to the crude RR (0.52), indicating that **occupation does not confound** the association between energy intake and CHD in this dataset.

## 2. Does the effect of high energy intake on CHD rate differ across job categories?

```
. stmh hieng, by(job)
```

```

failure _d:  chd
analysis time _t:  (dox-origin)/365.25
              origin:  time dob
enter on or after:  time doe

```

Maximum likelihood estimate of the rate ratio  
 comparing hieng==1 vs. hieng==0  
 by job

RR estimate, and lower and upper 95% confidence limits

job	RR	Lower	Upper
0	0.41	0.12	1.36
1	0.66	0.23	1.89
2	0.52	0.21	1.27

Overall estimate controlling for job

RR	chi2	P>chi2	[95% Conf. Interval]	
0.525	4.71	0.0299	0.291	0.949

```

Approx test for unequal RRs (effect modification):  chi2(2) =      0.33
                                                    Pr>chi2 =    0.8468

```

### Interpretation:

No evidence of effect modification (interaction) by occupation ( $p = 0.847$ ). The energy-CHD relationship is similar across job categories. This supports the stability of the main effect estimate.

The test for effect modification ( $p = 0.8468$ ) suggests no strong evidence that the relationship between energy intake and CHD varies across occupational categories. However, the similar pattern of protective effects across job categories (RRs ranging from 0.41 to 0.66) supports the validity of the overall adjusted estimate.

The adjusted estimated rate ratio (RR) for the effect of hieng on the rates of CHD is now 0.525, whereas the crude one was 0.520. Practically the adjustment did not affect the rate ratio. This appears to be because job does not affect the outcome.

**3. Investigate the relationship between high energy intake and CHD rate, controlling for both job and height group.**

```
. stmh hieng, by(job htgrp)
```

```
      failure _d:  chd
analysis time _t:  (dox-origin)/365.25
              origin:  time dob
enter on or after:  time doe
```

```
Maximum likelihood estimate of the rate ratio
comparing hieng==1 vs. hieng==0
by job htgrp
```

RR estimate, and lower and upper 95% confidence limits

job	htgrp	RR	Lower	Upper
0	[150-170cm)	0.69	0.10	4.88
0	[170-175cm)	0.41	0.04	3.95
0	[175-180cm)	0.22	0.02	2.15
1	[150-170cm)	0.59	0.15	2.37
1	[170-175cm)	1.09	0.18	6.52
2	[150-170cm)	0.41	0.08	2.26
2	[170-175cm)	0.91	0.18	4.51
2	[175-180cm)	0.50	0.11	2.22

Overall estimate controlling for job htgrp

RR	chi2	P>chi2	[95% Conf. Interval]	
0.569	3.48	0.0620	0.313	1.037

```
Approx test for unequal RRs (effect modification):  chi2(7) =      1.85
                                                    Pr>chi2 =    0.9677
```



Rate ratios stratified jointly by occupation (job) and height group (htgrp).

Overall Mantel–Haenszel RR (high vs. low energy), controlling for job and height:

$RR_{mh} = 0.569$  (95% CI: 0.313–1.037;  $p = 0.062$ ).

Test for effect modification (homogeneity of stratum-specific RRs):  $X^2(7) = 1.85$ ,  $p = 0.97$ .

**Interpretation:**

After adjustment for job and height, the inverse association persists but is **borderline** (CI includes 1). The non-significant homogeneity test suggests the association is **fairly consistent across strata**, supporting use of the common MH estimate. Nonetheless, variability and wide CIs (likely from sparse data in some strata) warrant caution; a regression approach (e.g., Poisson or Cox with interaction terms) would better assess potential interactions and confounding.