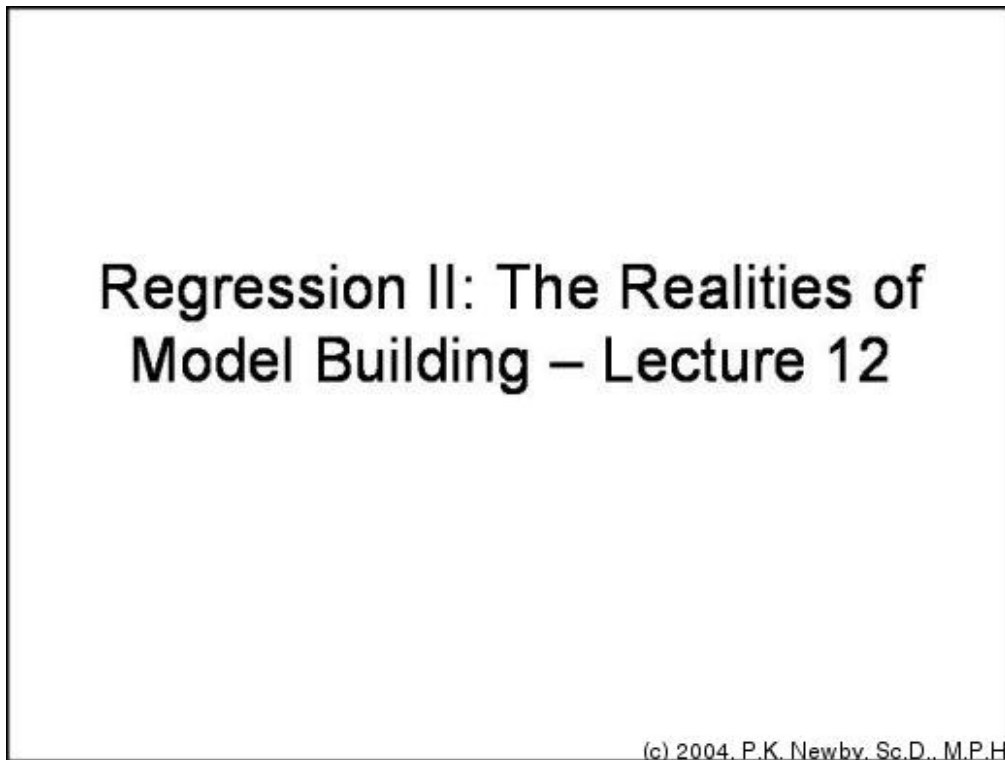
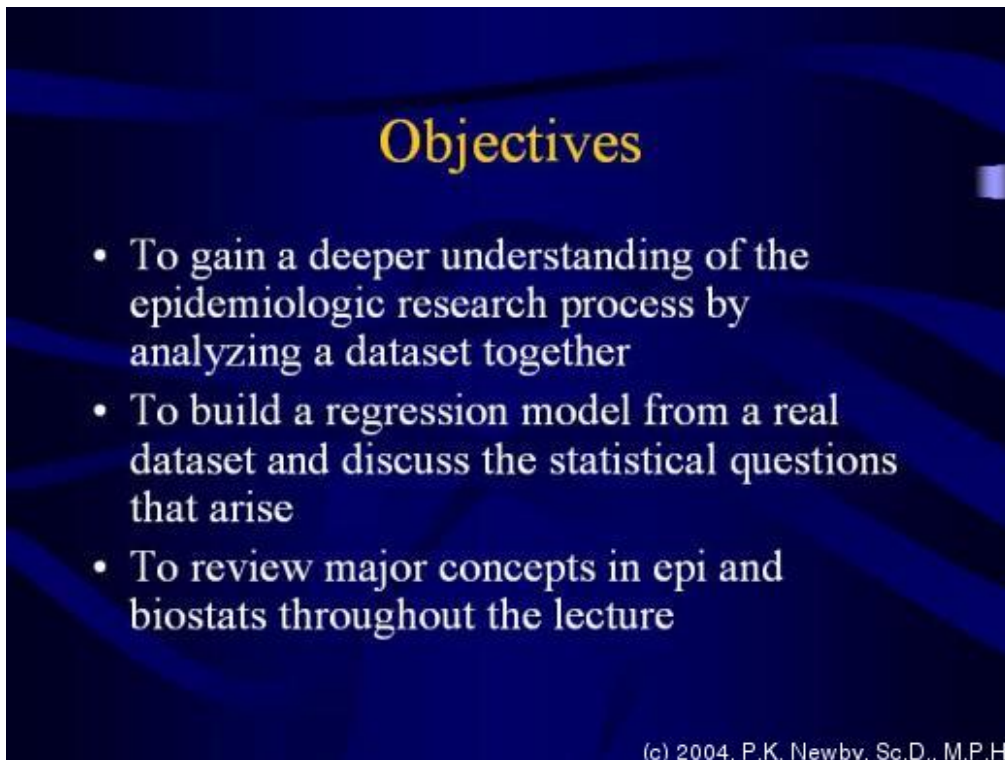


1. Lecture 12 - Introduction Slide



2. Objectives



3. Outline for lecture

## Outline for lecture

1. “Before you begin ... “
2. Pre-analysis phase
3. Analyzing a dataset
  - A. Descriptive statistics
  - B. Additional statistical tests
  - C. Building a regression model

(c) 2004, P.K. Newby, Sc.D., M.P.H.

4. Our example dataset

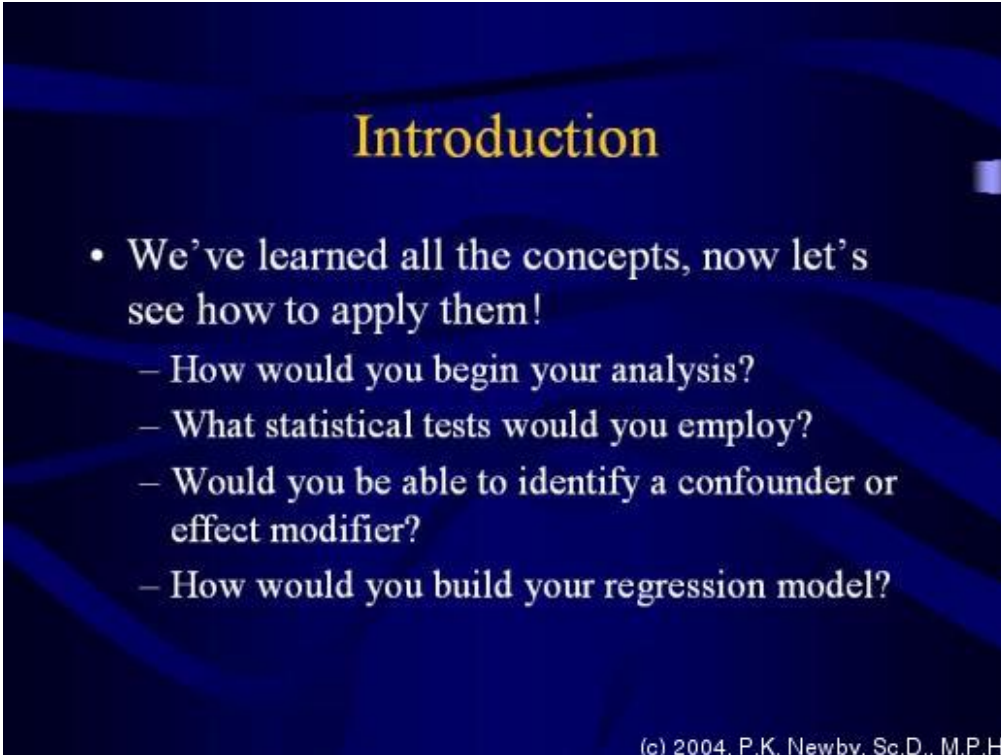
## Our example dataset

- Study design: Prospective cohort
- Study population: Men and women participating in the Baltimore Longitudinal Study of Aging
- Study aim: To examine the relation between diet and changes in body mass index (BMI) and waist circumference
- Hypothesis: A healthy dietary pattern will be related to smaller changes in BMI and waist circumference than other derived dietary patterns

[Newby et al. Am J Clin Nutr 2003;77(6):1417-1425]

(c) 2004, P.K. Newby, Sc.D., M.P.H.

5. Application of Concepts

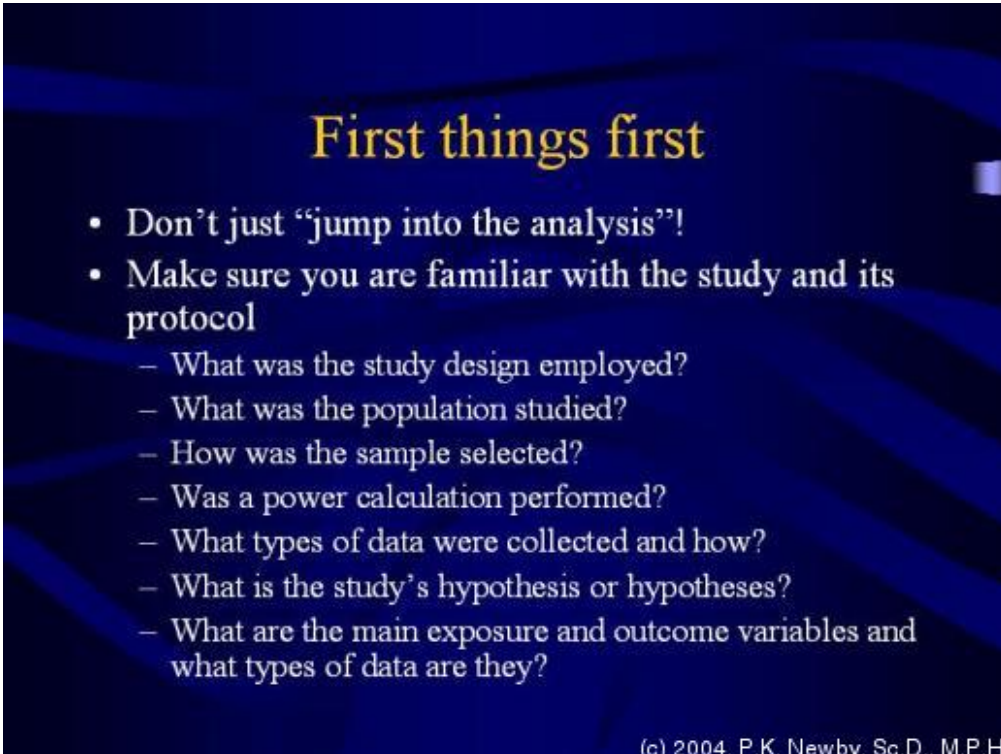


**Introduction**

- We've learned all the concepts, now let's see how to apply them!
  - How would you begin your analysis?
  - What statistical tests would you employ?
  - Would you be able to identify a confounder or effect modifier?
  - How would you build your regression model?

(c) 2004, P.K. Newby, Sc.D., M.P.H.

6. First things first

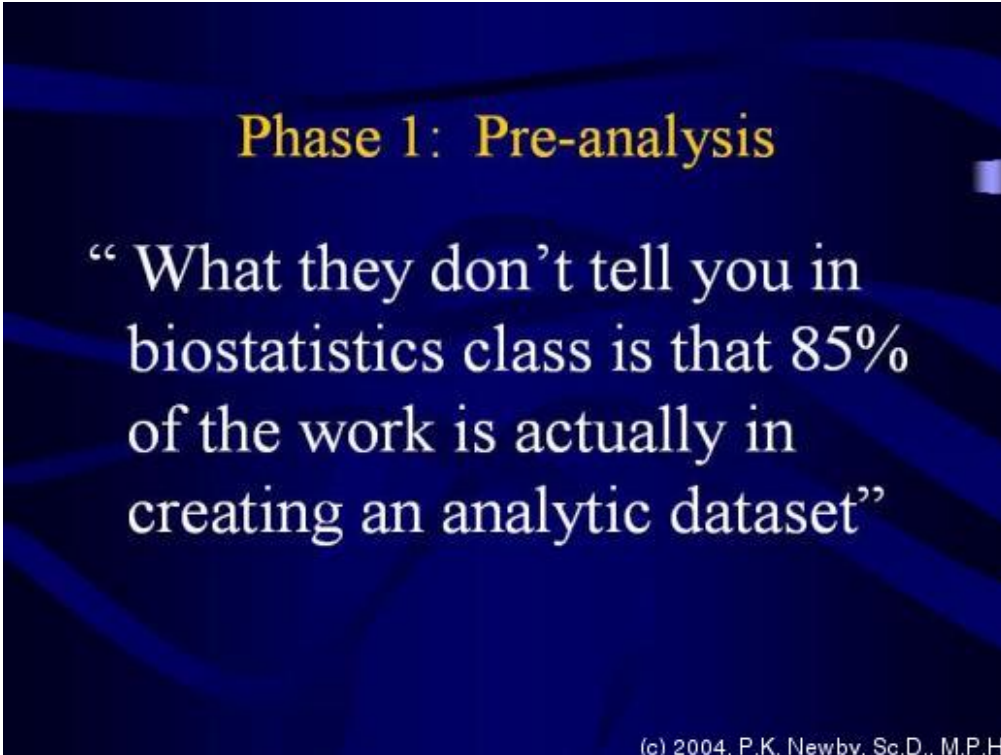


**First things first**

- Don't just "jump into the analysis"!
- Make sure you are familiar with the study and its protocol
  - What was the study design employed?
  - What was the population studied?
  - How was the sample selected?
  - Was a power calculation performed?
  - What types of data were collected and how?
  - What is the study's hypothesis or hypotheses?
  - What are the main exposure and outcome variables and what types of data are they?

(c) 2004, P.K. Newby, Sc.D., M.P.H.

7. Phase 1: Pre-analysis

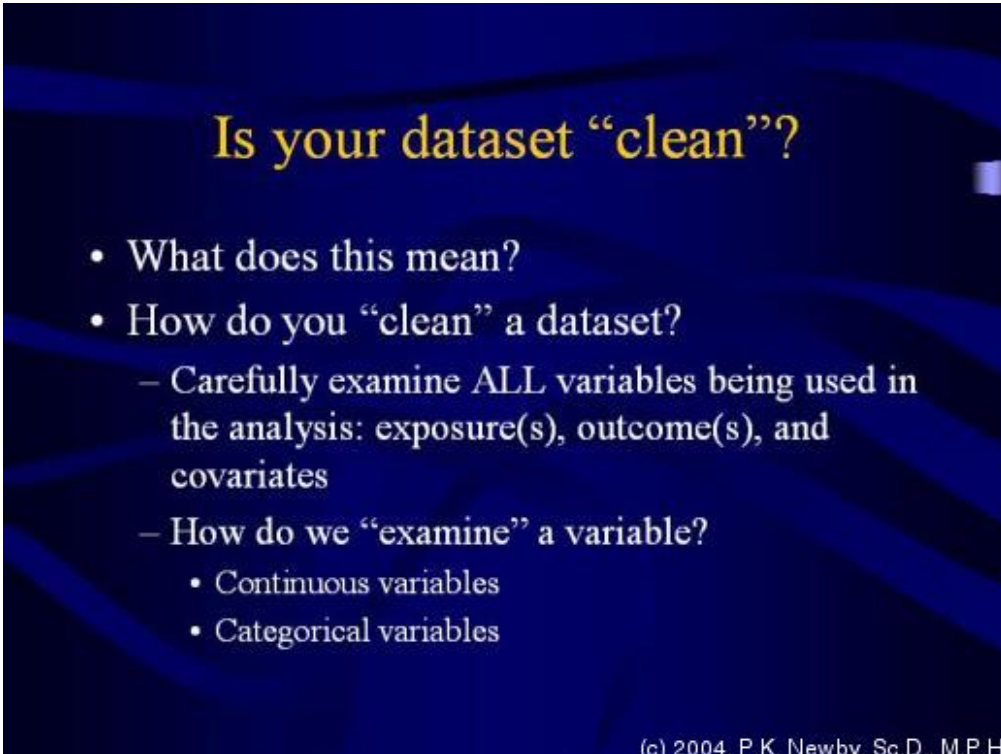


**Phase 1: Pre-analysis**

“ What they don’t tell you in biostatistics class is that 85% of the work is actually in creating an analytic dataset”

(c) 2004, P.K. Newby, Sc.D., M.P.H.

8. Is your dataset “clean”?



**Is your dataset “clean”?**

- What does this mean?
- How do you “clean” a dataset?
  - Carefully examine ALL variables being used in the analysis: exposure(s), outcome(s), and covariates
  - How do we “examine” a variable?
    - Continuous variables
    - Categorical variables

(c) 2004, P.K. Newby, Sc.D., M.P.H.

9. Data cleaning process

## Data cleaning process

1. Get a data coding book to see how your variables are recorded
2. Examine the values for each variable
  - Are values missing?
  - Are values biologically plausible?
  - Are values statistical outliers?
3. Look to the literature to see how other investigators have made decisions
4. Hire a good biostatistician!

(c) 2004, P.K. Newby, Sc.D., M.P.H.

10. Employ study restrictions

## Employ study restrictions

- Many (cohort) studies are conducted with numerous study aims, so any given study (analysis) will usually have its own, specific restrictions (exclusions) that uses a subset of the total dataset.
  - What do we mean by “study restrictions”
  - Why do studies have exclusions?
  - When else can study restrictions be employed?

(c) 2004, P.K. Newby, Sc.D., M.P.H.

11. Phase 2: Analyzing a dataset

**Phase 2: Analyzing a dataset**

- Step 1: Conduct descriptive statistics for Table 1
- Step 2: Conduct additional tests to describe the data
- Step 3: Build a regression model

Tables in an article are built based upon the above analyses.

(c) 2004, P.K. Newby, Sc.D., M.P.H.

12. Step 1: Conduct descriptive statistics

**Step 1:  
Conduct descriptive statistics**

- Why do we conduct and present descriptive statistics?
- What statistical tests do we use?
  - Continuous variables
  - Categorical variables

(c) 2004, P.K. Newby, Sc.D., M.P.H.

13. Example: Table 1 – Sample characteristics

**Example:  
Table 1 – Sample characteristics**

Sample characteristics	Women	Men
Number of subjects, N (%)	219 (47.7)	240 (52.3)
Age, y (mean ± SD)	57.3 ± 14.0	60.8 ± 13.3
BMI, kg/m <sup>2</sup> (mean ± SD)	24.7 ± 4.0	25.2 ± 3.0
Overweight, BMI 25-29.99, N (%)	68 (31.1)	102 (42.5)
Obese, BMI ≥ 30, N (%)	23 (10.5)	12 (5.0)
Waist circumference, cm (mean ± SD)	77.2 ± 9.4	90.4 ± 8.9
Physical activity, kcal/kg (mean ± SD)	15.2 ± 3.9	14.7 ± 3.8
Smoking status, N (%)		
Never smoker	168 (77.0)	181 (75.4)
Current smoker	23 (10.6)	22 (9.2)
Former smoker	27 (12.4)	37 (15.4)
Race/Ethnicity, N (%)		
White	206 (94.1)	231 (96.2)
African American	13 (5.9)	9 (3.8)
Vitamin users, N (%)	124 (56.6)	94 (39.2)

(c) 2004, P.K. Newby, Sc.D., M.P.H.

14. Review questions for Table 1

- Review questions for Table 1**
- Are there any notable results?
  - Does the above table tell us whether the characteristics are significantly different among men and women? What statistical test could we have performed to show this?
  - Would we use the same statistical test for continuous and nominal data?
- (c) 2004, P.K. Newby, Sc.D., M.P.H.

15. Step 2: Conduct additional tests to describe the data

## Step 2: Conduct additional tests to describe the data

- Depends on the study question!
- In this example, we presented sample characteristics (nutrient and demographic) by dietary pattern to explore whether there were differences by exposure group

(c) 2004, P.K. Newby, Sc.D., M.P.H.

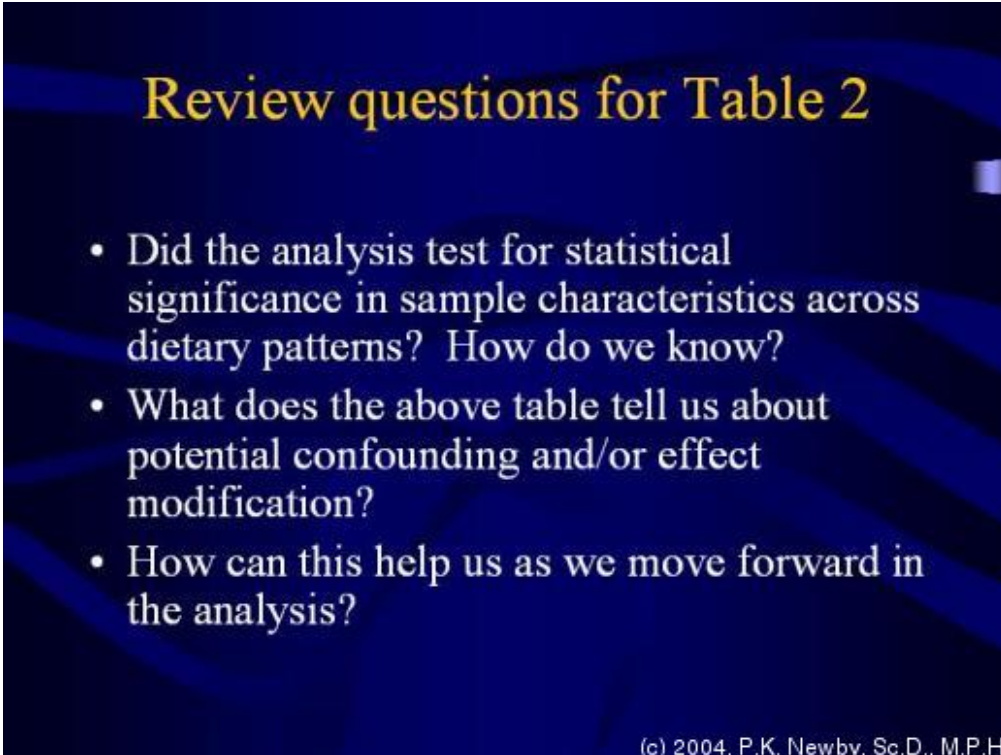
16. Example: Table 2 – Sample characteristics by exposure g...

## Example: Table 2 – Sample characteristics by exposure group (dietary pattern)

Sample characteristic	Cluster 1: White bread (n = 79)	Cluster 2: Alcohol (n = 60)	Cluster 3: Healthy (n = 98)	Cluster 4: Sweet (n = 140)	Cluster 5: Meat/potatoes (n = 82)	P value
Age, y	56.1 ± 1.5 <sup>a</sup>	58.2 ± 1.8 <sup>ab</sup>	63.4 ± 1.4 <sup>b</sup>	58.8 ± 1.1 <sup>ab</sup>	58.3 ± 1.5 <sup>ab</sup>	< 0.01
BMI, kg/m <sup>2</sup>	25.4 ± 0.4	25.2 ± 0.5	24.1 ± 0.4	25.1 ± 0.3	25.0 ± 0.4	NS
Waist circumference, cm	85.0 ± 1.0 <sup>ab</sup>	85.6 ± 1.1 <sup>a</sup>	81.6 ± 0.9 <sup>b</sup>	84.5 ± 0.7 <sup>ab</sup>	84.5 ± 1.0 <sup>ab</sup>	< 0.05
Physical activity, kcal/kg <sup>2</sup>	14.9 ± 0.4	14.9 ± 0.5	14.8 ± 0.4	15.1 ± 0.3	14.7 ± 0.4	NS
Female, N (%)	34 (43.0) <sup>ab</sup>	20 (33.3) <sup>a</sup>	57 (58.2) <sup>b</sup>	62 (44.0) <sup>ab</sup>	46 (56.1) <sup>ab</sup>	< 0.05
Vitamin users, N (%)	38 (48.1) <sup>ab</sup>	34 (56.7) <sup>ab</sup>	61 (62.2) <sup>a</sup>	53 (37.6) <sup>b</sup>	32 (39.0) <sup>b</sup>	< 0.001
Smoking status, N (%)						< 0.01
Current smoker	8 (10.1) <sup>ab</sup>	14 (29.3) <sup>a</sup>	4 (4.1) <sup>b</sup>	11 (7.8) <sup>b</sup>	8 (9.8) <sup>b</sup>	< 0.01
Never smoker	60 (76.0) <sup>a</sup>	33 (55.0) <sup>b</sup>	81 (82.6) <sup>a</sup>	111 (78.7) <sup>a</sup>	65 (80.3) <sup>a</sup>	< 0.01
Former smoker	11 (13.9)	13 (21.7)	13 (13.3)	19 (13.5)	8 (9.9)	NS

(c) 2004, P.K. Newby, Sc.D., M.P.H.

17. Review questions for Table 2

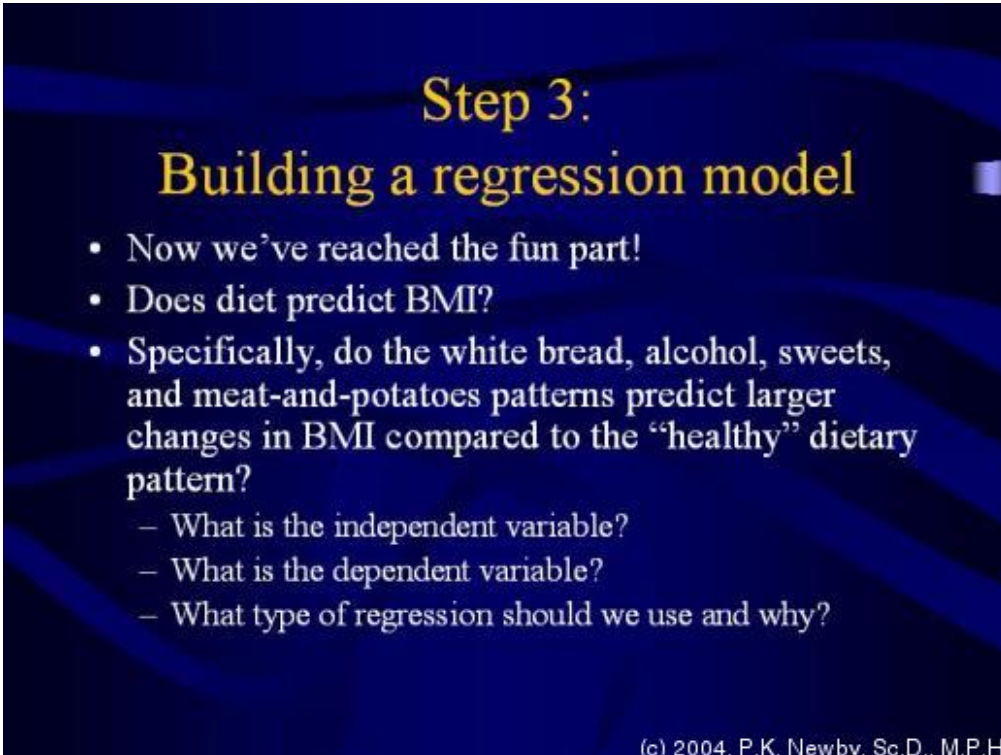


**Review questions for Table 2**

- Did the analysis test for statistical significance in sample characteristics across dietary patterns? How do we know?
- What does the above table tell us about potential confounding and/or effect modification?
- How can this help us as we move forward in the analysis?

(c) 2004, P.K. Newby, Sc.D., M.P.H.

18. Step 3: Building a regression model

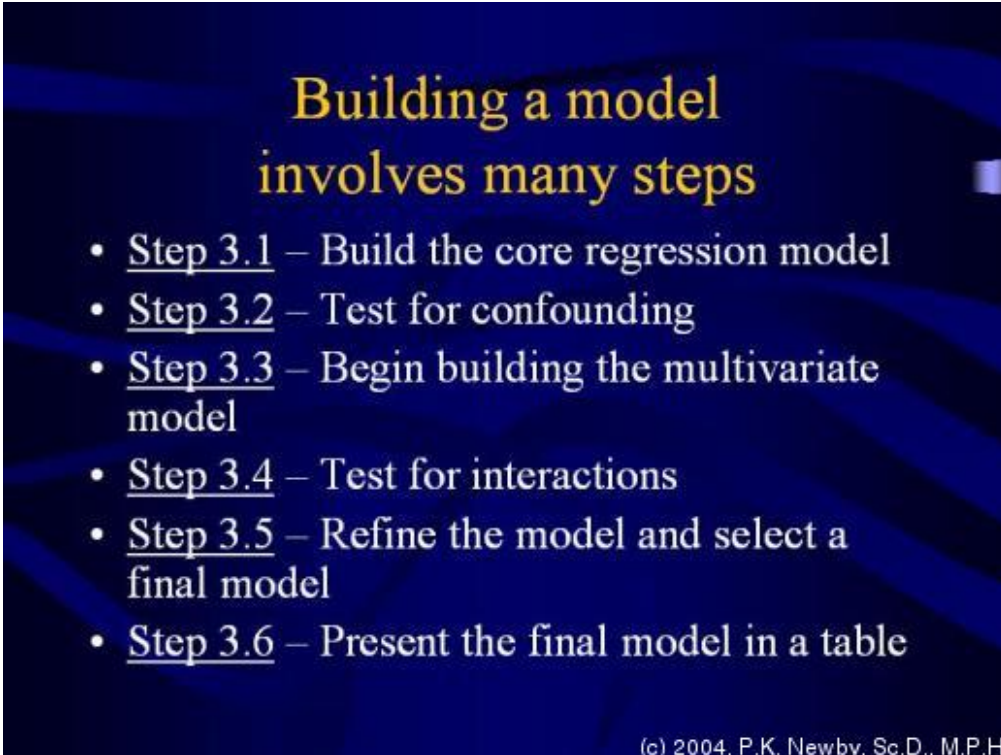


**Step 3:  
Building a regression model**

- Now we've reached the fun part!
- Does diet predict BMI?
- Specifically, do the white bread, alcohol, sweets, and meat-and-potatoes patterns predict larger changes in BMI compared to the "healthy" dietary pattern?
  - What is the independent variable?
  - What is the dependent variable?
  - What type of regression should we use and why?

(c) 2004, P.K. Newby, Sc.D., M.P.H.

19. Building a model involves many steps

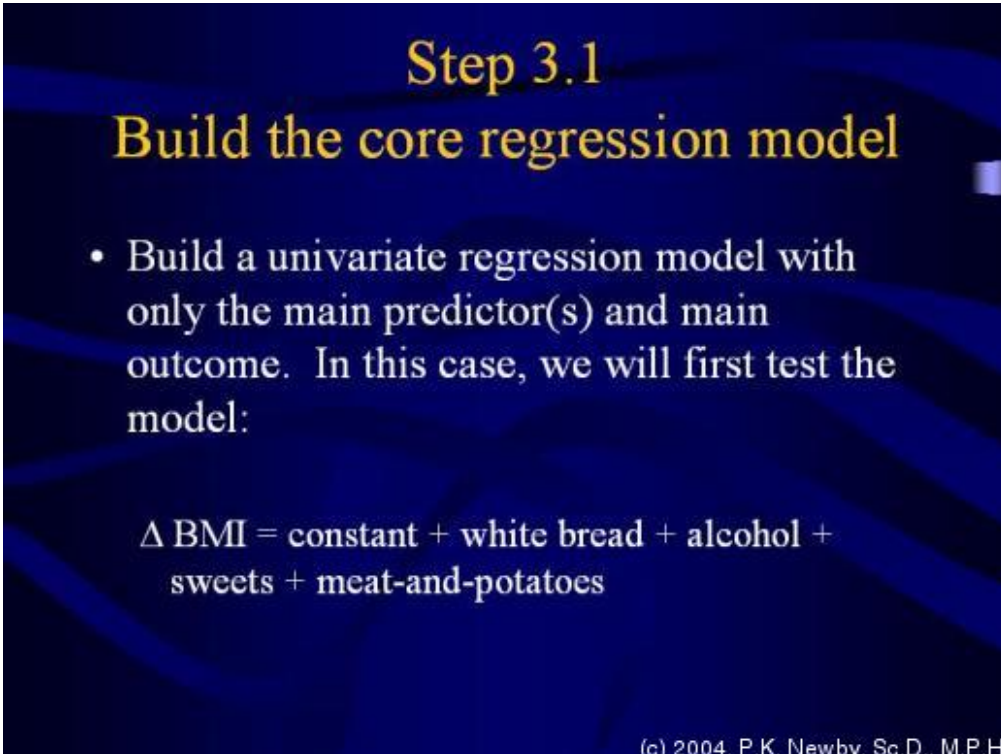


**Building a model involves many steps**

- Step 3.1 – Build the core regression model
- Step 3.2 – Test for confounding
- Step 3.3 – Begin building the multivariate model
- Step 3.4 – Test for interactions
- Step 3.5 – Refine the model and select a final model
- Step 3.6 – Present the final model in a table

(c) 2004, P.K. Newby, Sc.D., M.P.H.

20. Step 3.1: Build the core regression model



**Step 3.1**  
**Build the core regression model**

- Build a univariate regression model with only the main predictor(s) and main outcome. In this case, we will first test the model:

$$\Delta \text{BMI} = \text{constant} + \text{white bread} + \text{alcohol} + \text{sweets} + \text{meat-and-potatoes}$$

(c) 2004, P.K. Newby, Sc.D., M.P.H.

21. Univariate regression results

**Univariate regression results**

You will need to refer to the univariate regression results when determining whether or not confounding occurred.

- White bread:  $\beta = 0.08$
- Alcohol:  $\beta = 0.08$
- Sweets:  $\beta = 0.03$
- Meat-and-potatoes:  $\beta = 0.26$

(c) 2004, P.K. Newby, Sc.D., M.P.H.

22. Step 3.2 Testing for confounding

**Step 3.2**  
**Testing for confounding**

- Descriptive analyses suggest potential confounding, since there were differences in sample characteristics by exposure status
- We still don't know whether those variables are related to our outcome
- Perform univariate regression analysis with each potential confounder to see if they predict our outcome,  $\Delta$  BMI. Examples:
  - $\Delta$  BMI = constant + age
  - $\Delta$  BMI = constant + sex
  - $\Delta$  BMI = constant + baseline BMI

(c) 2004, P.K. Newby, Sc.D., M.P.H.

23. Review questions - confounding

**Review questions - confounding**

- What is the definition of a confounder?
- Regression is one way of testing whether two variables are related, how can we tell if the relationship is significant?
- Are there are other ways of examining whether variables are associated, other than regression analysis?

(c) 2004, P.K. Newby, Sc.D., M.P.H.

24. Step 3.3: Begin building the multivariate model

**Step 3.3:  
Begin building  
the multivariate model**

- Start with the core model
- Add potential confounders, one at a time, to see if the regression coefficients change
- We aren't really looking at the coefficient of the added variable, just our main exposures
- Example:  
$$\Delta \text{BMI} = \text{constant} + \text{white bread} + \text{alcohol} + \text{sweets} + \text{meat-and-potatoes} + \text{age}$$

(c) 2004, P.K. Newby, Sc.D., M.P.H.

25. Example: core model plus age

**Example: core model plus age**

<u>Univariate:</u>	<u>Adjusted for age:</u>
• White bread: $\beta = 0.08$	• White bread: $\beta = 0.06$
• Alcohol: $\beta = 0.08$	• Alcohol: $\beta = 0.07$
• Sweets: $\beta = 0.03$	• Sweets: $\beta = 0.01$
• Meat and potatoes: $\beta = 0.25$	• Meat and potatoes: $\beta = 0.23$

(c) 2004, P.K. Newby, Sc.D., M.P.H.

26. Continue building the model (1)

**Continue building the model (1)**

- Continue building the model by adding individual variables one at a time and seeing if the regression coefficients change (10% rule)
- Do NOT look to  $P$  values during this process
- It is not uncommon to have a significant predictor variable lose significance once additional variables are added to the model, often because the predictor was confounded
  - Is there another reason why a variable that was significant in a smaller model may lose significance in a larger, multivariate model?

(c) 2004, P.K. Newby, Sc.D., M.P.H.

27. Continue building the model (2)

**Continue building the model (2)**

- Note that checking for confounding in regression repeats some of our earlier work, which may have shown significant associations, but adding potential confounders one at a time allows us to see how much confounding was occurring
- Many investigators choose to keep variables in the model that do not appear to be confounding based upon their statistical tests but are “known” confounders

Review question: when should a variable NOT be added to the model?

(c) 2004, P.K. Newby, Sc.D., M.P.H.

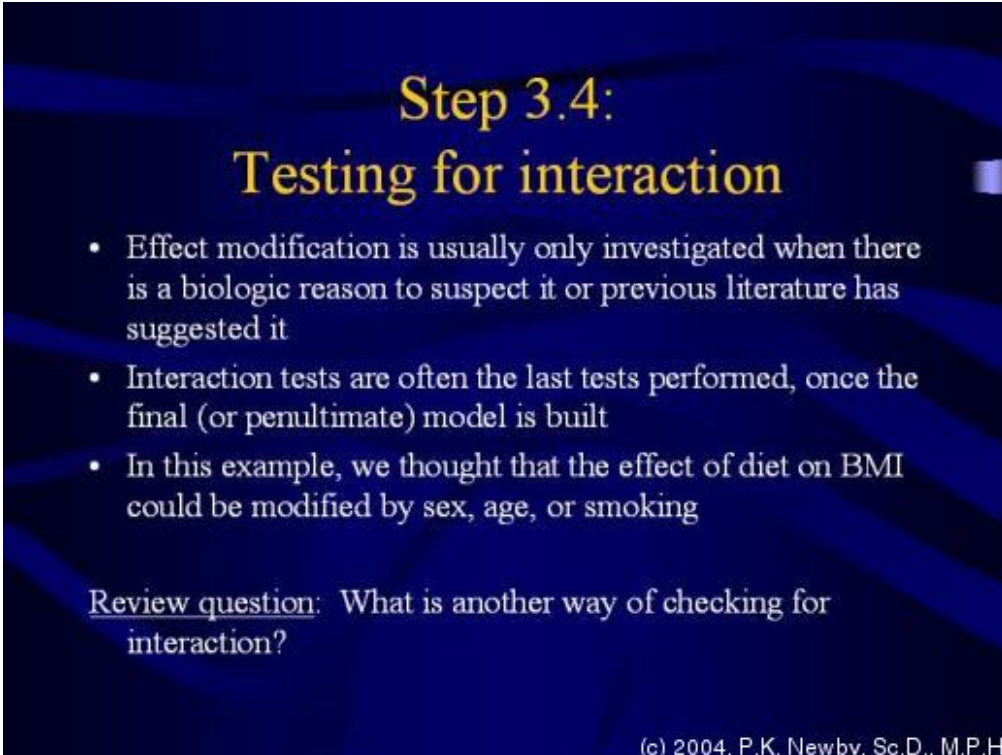
28. Example: core model versus multivariate adjusted model

**Example: core model versus multivariate adjusted model**

<u>Univariate:</u>	<u>Multivariate adjusted:</u>
• White bread: $\beta = 0.08$	• White bread: $\beta = 0.04$
• Alcohol: $\beta = 0.08$	• Alcohol: $\beta = 0.05$
• Sweets: $\beta = 0.03$	• Sweets: $\beta = 0.02$
• Meat and potatoes: $\beta = 0.25$	• Meat and potatoes: $\beta = 0.25$

(c) 2004, P.K. Newby, Sc.D., M.P.H.

29. Step 3.4: Testing for interaction



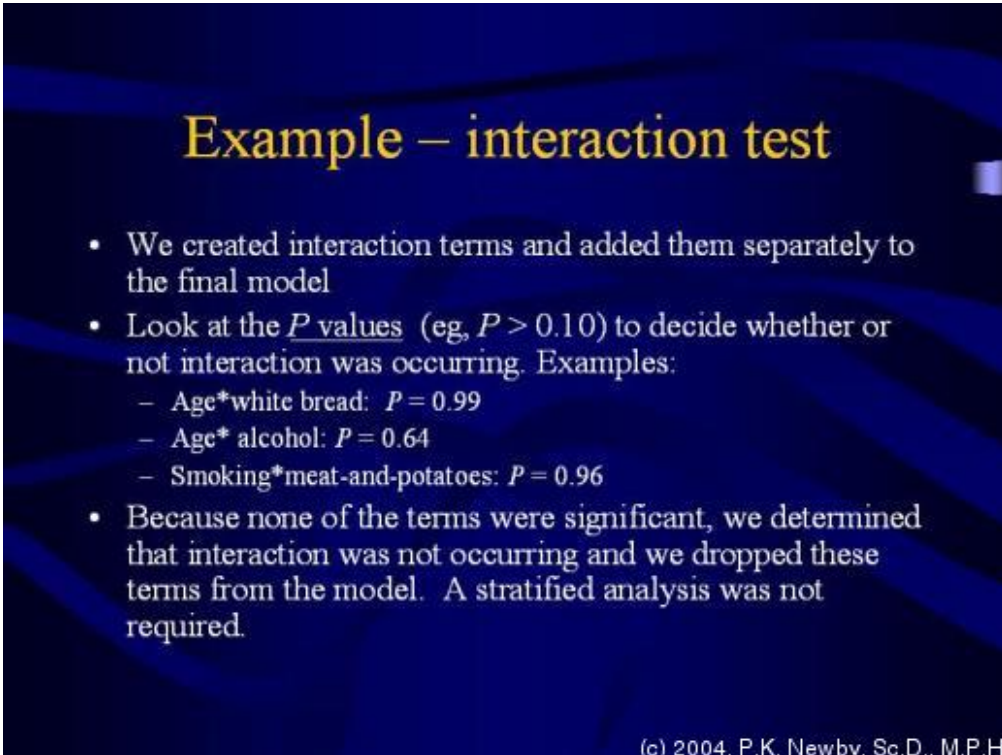
**Step 3.4:  
Testing for interaction**

- Effect modification is usually only investigated when there is a biologic reason to suspect it or previous literature has suggested it
- Interaction tests are often the last tests performed, once the final (or penultimate) model is built
- In this example, we thought that the effect of diet on BMI could be modified by sex, age, or smoking

Review question: What is another way of checking for interaction?

(c) 2004, P.K. Newby, Sc.D., M.P.H.

30. Example – interaction test

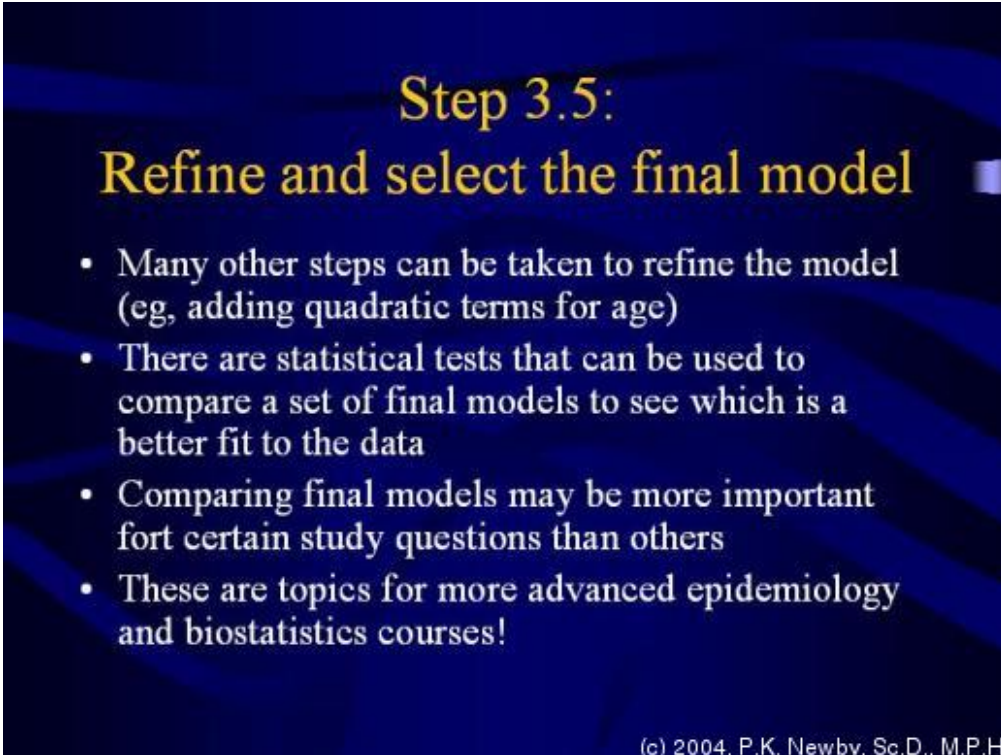


**Example – interaction test**

- We created interaction terms and added them separately to the final model
- Look at the P values (eg,  $P > 0.10$ ) to decide whether or not interaction was occurring. Examples:
  - Age\*white bread:  $P = 0.99$
  - Age\* alcohol:  $P = 0.64$
  - Smoking\*meat-and-potatoes:  $P = 0.96$
- Because none of the terms were significant, we determined that interaction was not occurring and we dropped these terms from the model. A stratified analysis was not required.

(c) 2004, P.K. Newby, Sc.D., M.P.H.

31. Step 3.5: Refine and select the final model

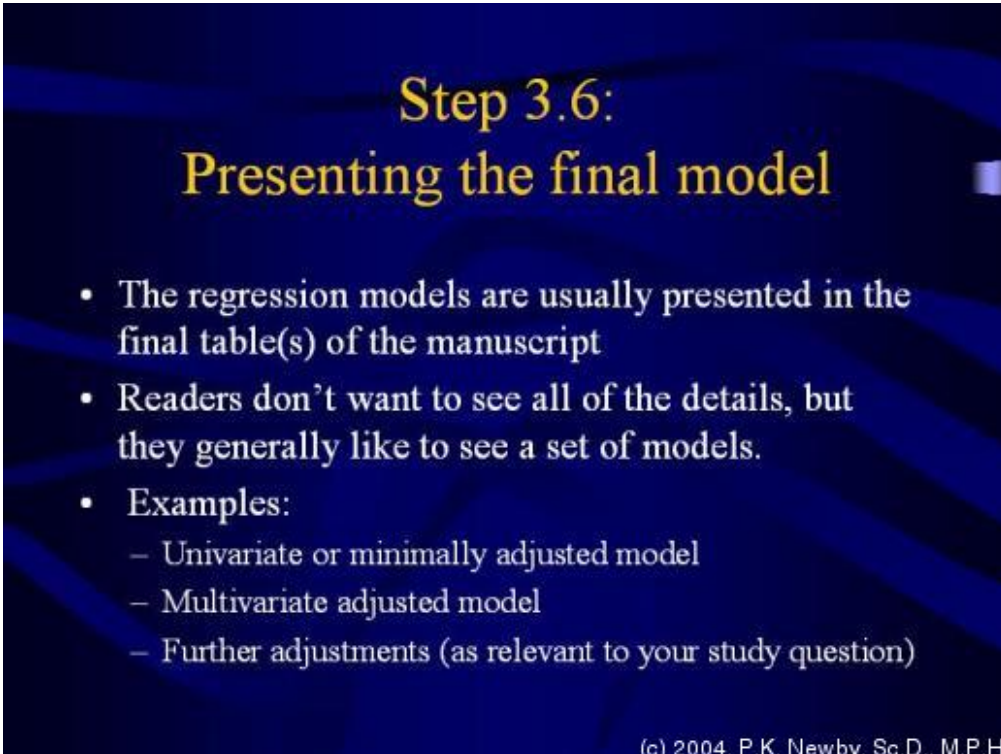


**Step 3.5:**  
**Refine and select the final model**

- Many other steps can be taken to refine the model (eg, adding quadratic terms for age)
- There are statistical tests that can be used to compare a set of final models to see which is a better fit to the data
- Comparing final models may be more important for certain study questions than others
- These are topics for more advanced epidemiology and biostatistics courses!

(c) 2004, P.K. Newby, Sc.D., M.P.H.

32. Step 3.6: Presenting the final model



**Step 3.6:**  
**Presenting the final model**

- The regression models are usually presented in the final table(s) of the manuscript
- Readers don't want to see all of the details, but they generally like to see a set of models.
- Examples:
  - Univariate or minimally adjusted model
  - Multivariate adjusted model
  - Further adjustments (as relevant to your study question)

(c) 2004, P.K. Newby, Sc.D., M.P.H.

33. Example: Table 4 - Regression models

### Example: Table 4 - Regression models

Dietary pattern	Change in BMI β (SE)
<b>Cluster 2: White bread</b>	
Adjusted for age and sex	0.06 (0.09)
Multivariate adjusted <sup>4</sup>	0.05 (0.08)
Multivariate adjusted+energy	0.05 (0.09)
<b>Cluster 3: Alcohol</b>	
Adjusted for age and sex	0.07 (0.09)
Multivariate adjusted <sup>4</sup>	0.06 (0.10)
Multivariate adjusted+energy	0.06 (0.10)
<b>Cluster 4: Sweet</b>	
Adjusted for age and sex	0.01 (0.09)
Multivariate adjusted <sup>4</sup>	0.02 (0.08)
Multivariate adjusted+energy	0.04 (0.08)
<b>Cluster 5: Meat/potatoes</b>	
Adjusted for age and sex	0.23 (0.09) <sup>5</sup>
Multivariate adjusted <sup>4</sup>	0.25 (0.09) <sup>5</sup>
Multivariate adjusted+energy	0.26 (0.09) <sup>5</sup>

(c) 2004, P.K. Newby, Sc.D., M.P.H.

34. Final models in our example

### Final models in our example

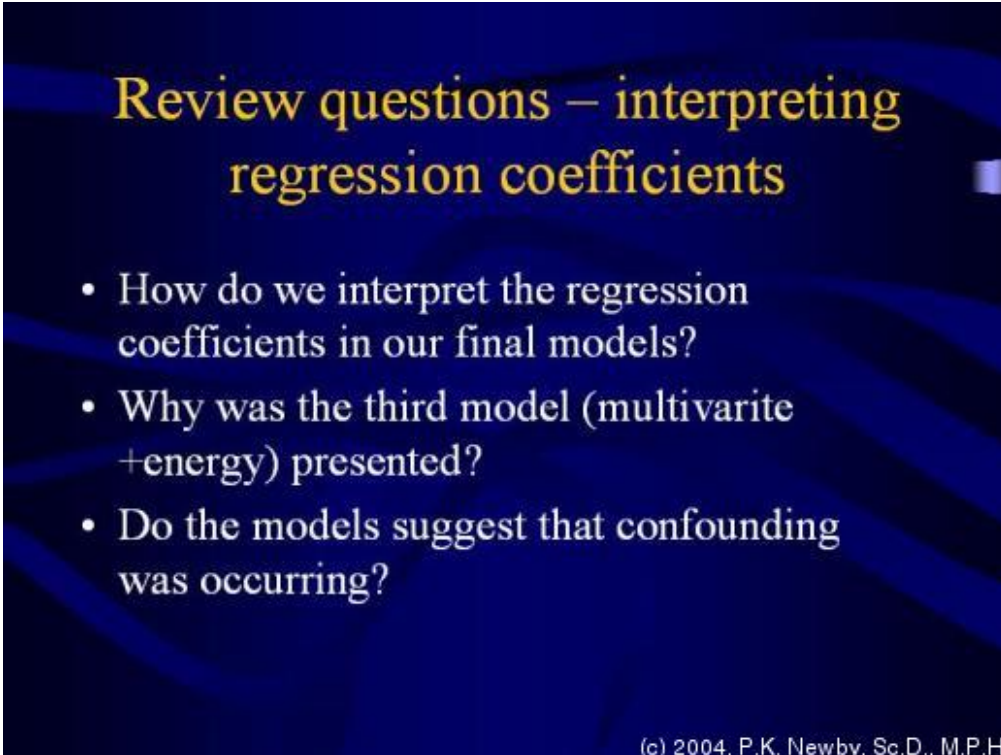
**Model 1: minimally adjusted**  
 $\Delta \text{ BMI} = \text{constant} + \text{white bread} + \text{alcohol} + \text{sweets} + \text{meat-and-potato} + \text{age} + \text{sex} + \text{baseline BMI}$

**Model 2: multivariate adjusted**  
 $\Delta \text{ BMI} = \text{constant} + \text{white bread} + \text{alcohol} + \text{sweets} + \text{meat-and-potato} + \text{age} + \text{sex} + \text{baseline BMI} + \text{ethnicity} + \text{physical activity} + \text{past smoking} + \text{current smoking} + \text{vitamin supplement use}$

**Model 3: further adjusted**  
 $\Delta \text{ BMI} = \text{constant} + \text{white bread} + \text{alcohol} + \text{sweets} + \text{meat-and-potato} + \text{age} + \text{sex} + \text{baseline BMI} + \text{ethnicity} + \text{physical activity} + \text{past smoking} + \text{current smoking} + \text{vitamin supplement use} + \text{energy}$

(c) 2004, P.K. Newby, Sc.D., M.P.H.

35. Review questions – interpreting regression coefficient...

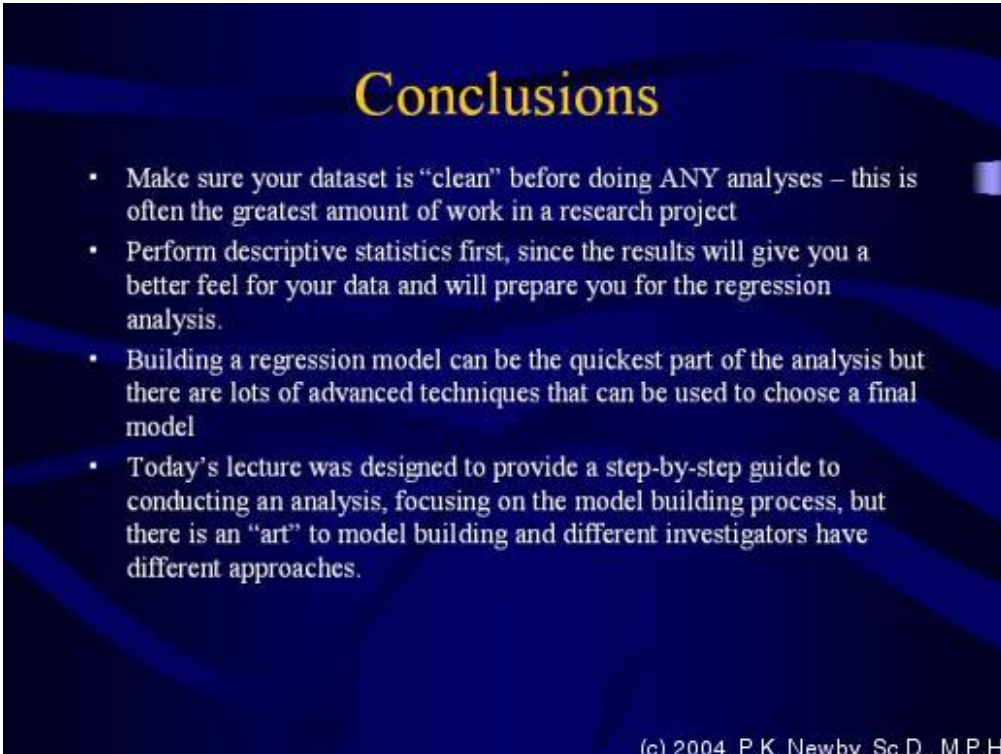


**Review questions – interpreting regression coefficients**

- How do we interpret the regression coefficients in our final models?
- Why was the third model (multivariate +energy) presented?
- Do the models suggest that confounding was occurring?

(c) 2004, P.K. Newby, Sc.D., M.P.H.

36. Conclusions



**Conclusions**

- Make sure your dataset is “clean” before doing ANY analyses – this is often the greatest amount of work in a research project
- Perform descriptive statistics first, since the results will give you a better feel for your data and will prepare you for the regression analysis.
- Building a regression model can be the quickest part of the analysis but there are lots of advanced techniques that can be used to choose a final model
- Today’s lecture was designed to provide a step-by-step guide to conducting an analysis, focusing on the model building process, but there is an “art” to model building and different investigators have different approaches.

(c) 2004, P.K. Newby, Sc.D., M.P.H.