Statistics in Healthcare

Unit 9:

Overview/Teasers

Overview

- Regression II: Logistic regression; Cox regression

Common statistics for various types of outcome data

Outcome Variable	Are the observations independent or correlated?		Alternatives (assumptions violated)
	independent	correlated	
Continuous (e.g. pain scale, cognitive function)	Ttest ANOVA Linear correlation Linear regression	Paired ttest Repeated-measures ANOVA Mixed models/GEE modeling	Wilcoxon sign-rank test Wilcoxon rank-sum test Kruskal-Wallis test Spearman rank correlation coefficient
Binary or categorical (e.g. fracture yes/no)	Risk difference/Relative risks Chi-square test Logistic regression	McNemar's test Conditional logistic regression GEE modeling	Fisher's exact test McNemar's exact test
Time-to-event (e.g. time to fracture)	Rate ratio Kaplan-Meier statistics Cox regression	Frailty model (beyond the scope of this course)	Time-varying effects (beyond the scope of this course)

Teaser 1, Unit 9

- 2009 headline (NBCnews.com):

  Eating a lot of red meat may up mortality risk

  Study's findings support advice to cut intake to reduce cancer, heart disease.

  "The largest study of its kind finds that older Americans who eat large amounts of red meat and processed meats face a greater risk of death from heart disease and cancer."

Risk factors cluster!

Table 1. Selected Age-Adjusted Characteristics of the National Institutes of Health–AARP Cohort by Red Meat Quintile Category[a]

Characteristic	Red Meat Intake Quintile, g/1000 kcal				
	Q1	Q2	Q3	Q4	Q5
Men (n=322 263)					
Meat intake					
Red meat, g/1000 kcal	9.3	21.4	31.5	43.1	68.1
White meat, g/1000 kcal	36.6	32.2	30.7	30.4	30.9
Processed meat, g/1000 kcal	5.1	7.8	10.3	13.3	19.4
Age, y	62.8	62.8	62.5	62.3	61.7
Race, %					
Non-Hispanic white	88.6	91.8	93.1	94.0	94.1
Non-Hispanic black	4.2	3.2	2.7	2.2	1.9
Hispanic/Asian/Pacific Islander/American Indian/Alaskan native/unknown	7.2	5.0	4.2	3.8	4.0
Positive family history of cancer,%	47.0	47.7	48.4	48.6	47.8
Currently married, %	80.8	84.4	86.1	86.7	85.6
BMI	25.9	26.7	27.1	27.6	28.3
Smoking history, %[b]					
Never smoker	34.4	30.5	28.8	27.6	25.4
Former smoker	56.5	58.1	57.5	57.1	55.8
Current smoker or having quit <1 y prior	4.9	7.6	9.9	11.4	14.8
Education, college graduate or postgraduate, %	53.0	47.3	45.1	42.3	39.1
Vigorous physical activity ≥5 times/wk, %	30.7	23.6	20.5	18.6	16.3
Dietary intake					
Energy, kcal/d	1899	1955	1998	2038	2116
Fruit, servings/1000 kcal	2.3	1.8	1.6	1.4	1.1
Vegetables, servings/1000 kcal	2.4	2.1	2.0	2.0	1.9

Reproduced with permission from Table 1 of: Sinha R, Cross AJ, Graubard BI, Leitzmann MF, Schatzkin A. Meat intake and mortality: a prospective study of over half a million people. *Arch Intern Med* 2009;169:562-71.

Statistics in Medicine

Module 1:

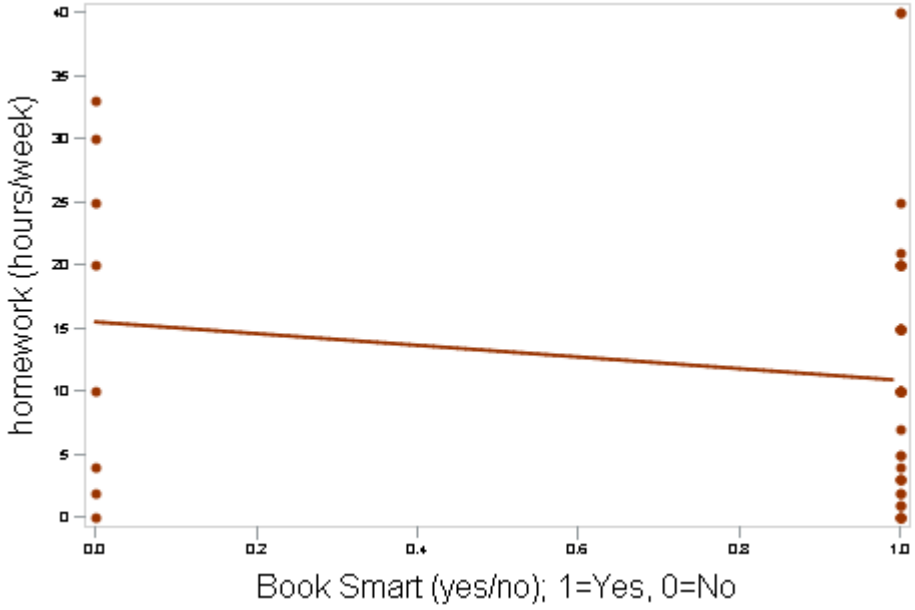
Logistic regression

Binary or categorical outcomes (proportions)

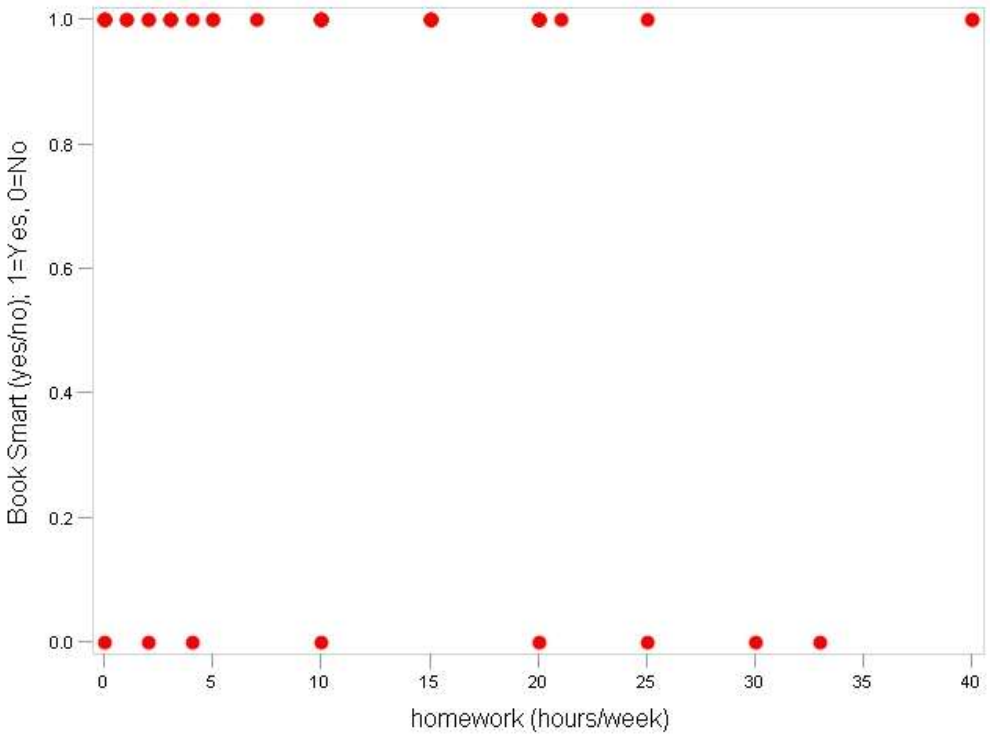
Outcome Variable	**Are the observations correlated?**		**Alternatives if sparse data:**
	independent	correlated	
Binary or categorical (e.g. fracture, yes/no)	**Risk difference/relative risks** (2x2 table) **Chi-square test** (RxC table) **Logistic regression** (multivariate regression technique)	**McNemar's chi-square test** (2x2 table) **Conditional logistic regression** (multivariate regression technique) **GEE modeling** (multivariate regression technique)	**McNemar's exact test** (alternative to McNemar's chi-square, for sparse data) **Fisher's exact test** (alternative to the chi-square, for sparse data)

Recall: linear regression with a binary predictor!

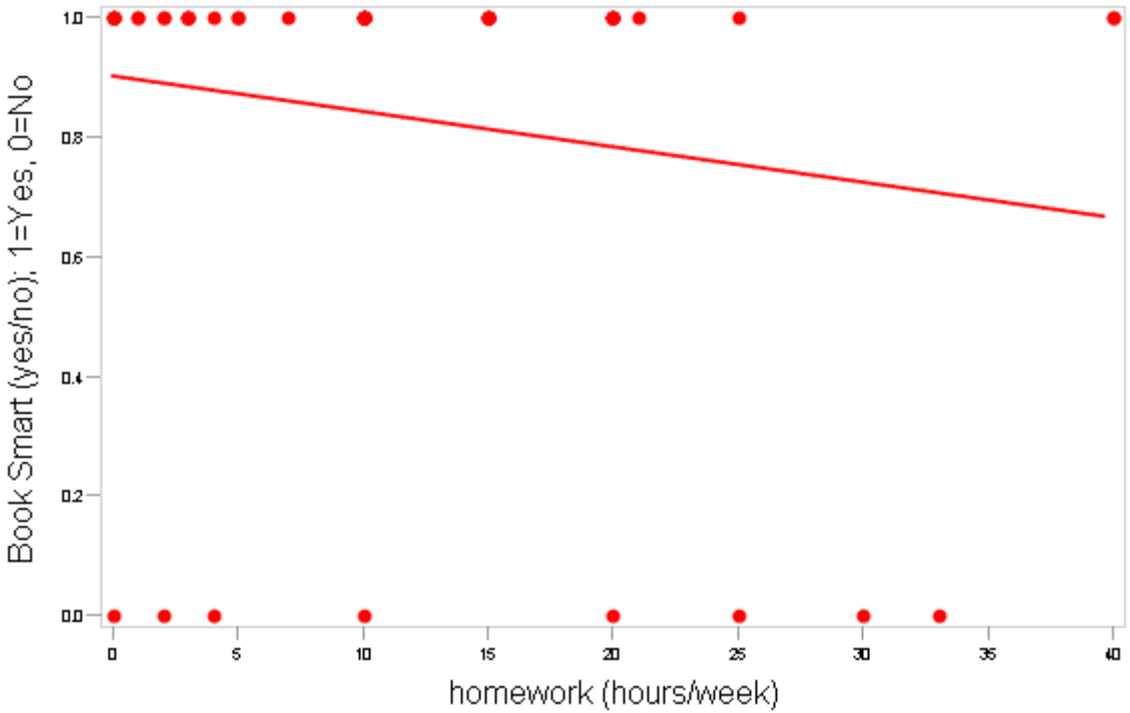
From our example dataset. Do those who think they are "book smart" spend more time on homework than those who think they are "street smart"?



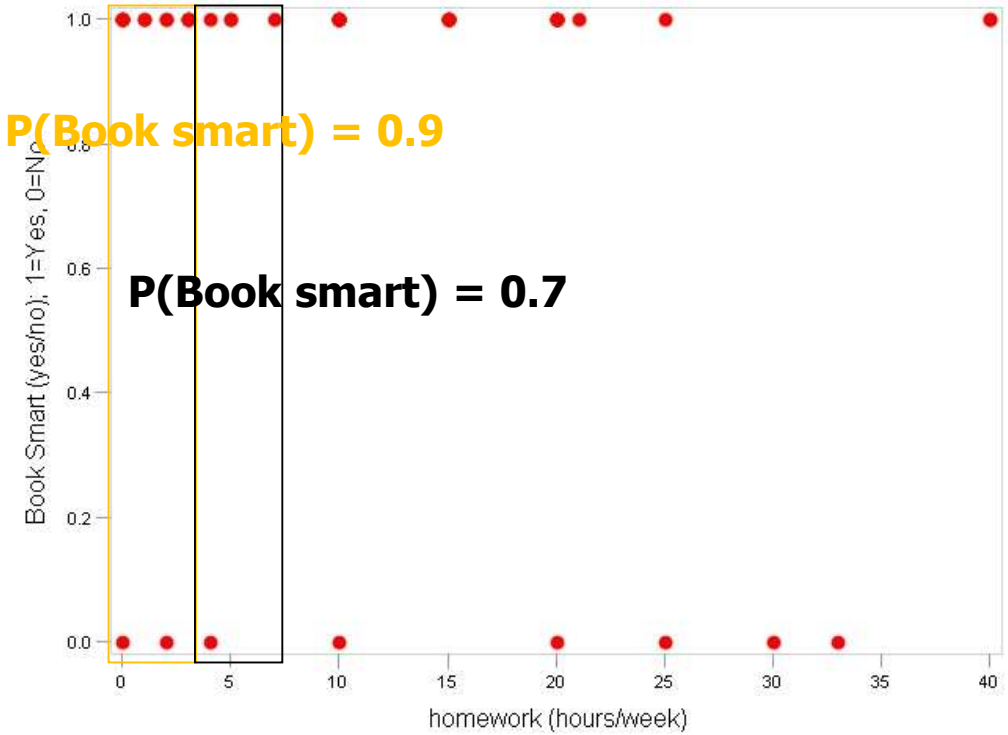
Flip x and y; now the outcome is binary…



Is a line a good fit for these data? Not so much!



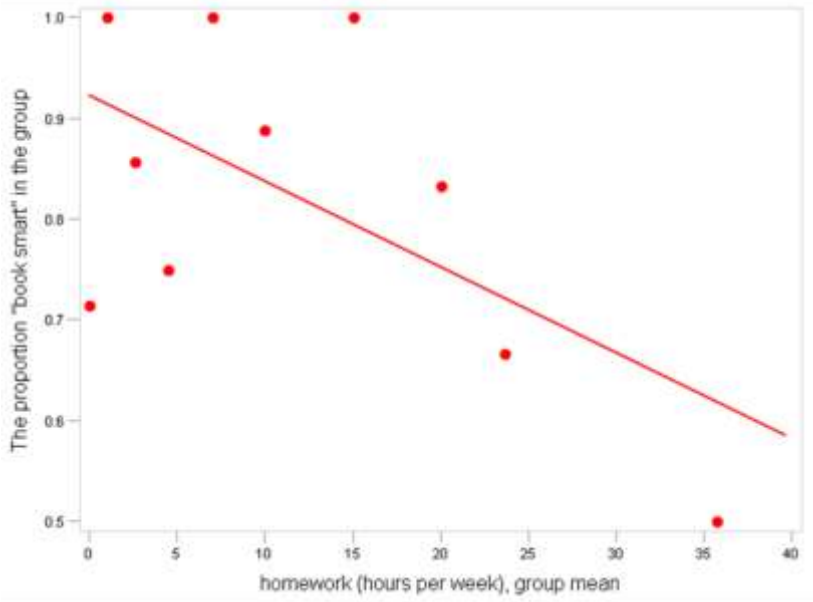
How could we transform the outcome variable?



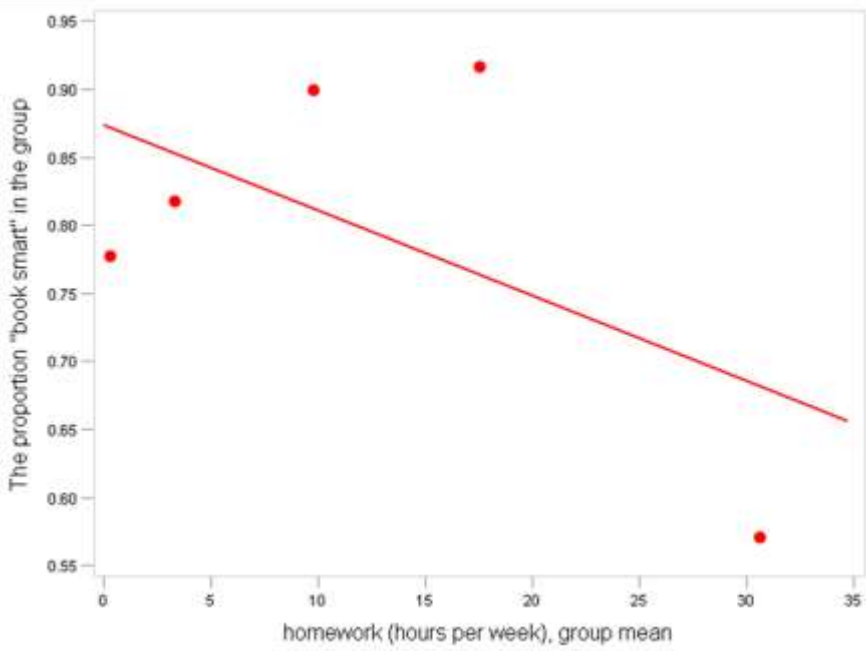
Model as Probabilities

What if we made the outcome variable a probability instead?

10 groups



5 groups



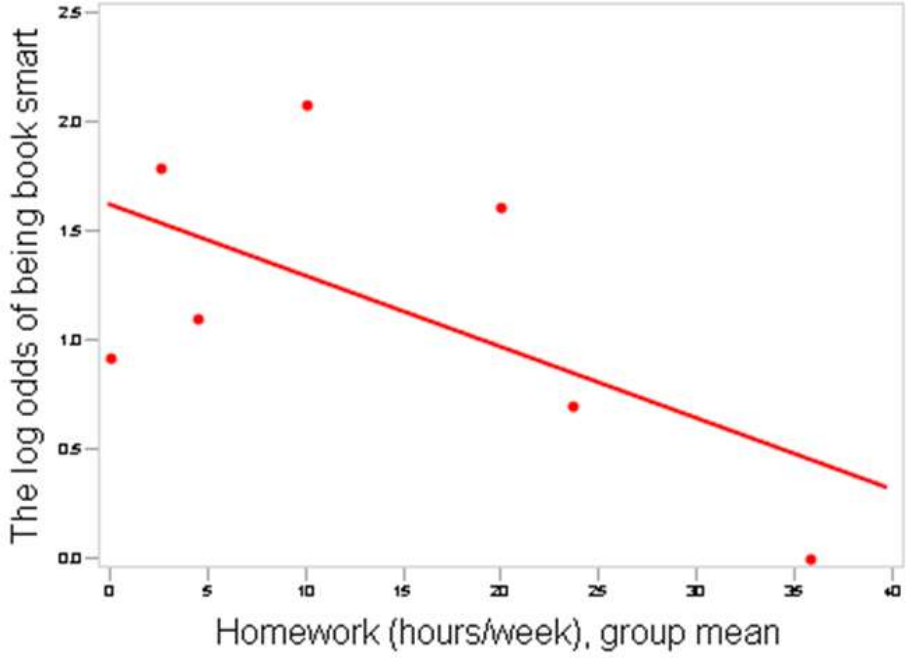
Mathematically better: the logit of the outcome!

- Probability p: [0 ,1]
- However, intercept of model (p) can go to less than 0, or more than 1

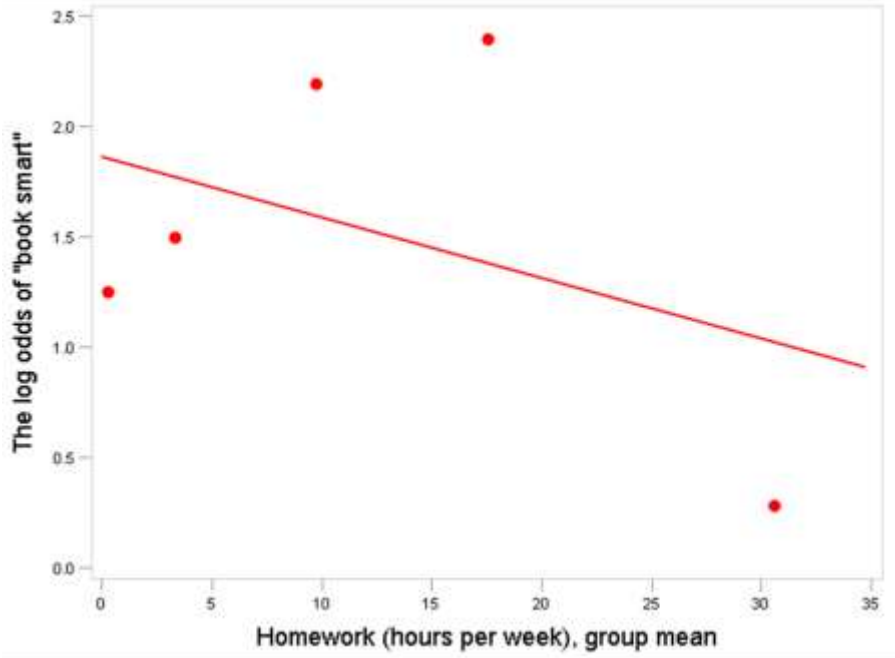
- Therefore, use odds = p/1-p [0, +inf]
- Logit = ln(p/1-p) [-inf, +inf]

The logit of the outcome:

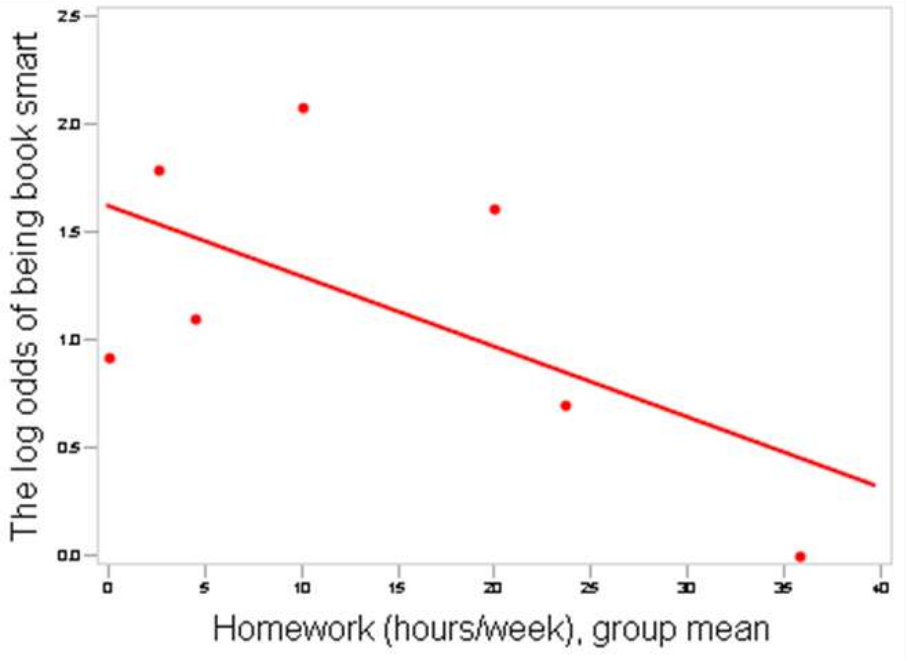
10 groups

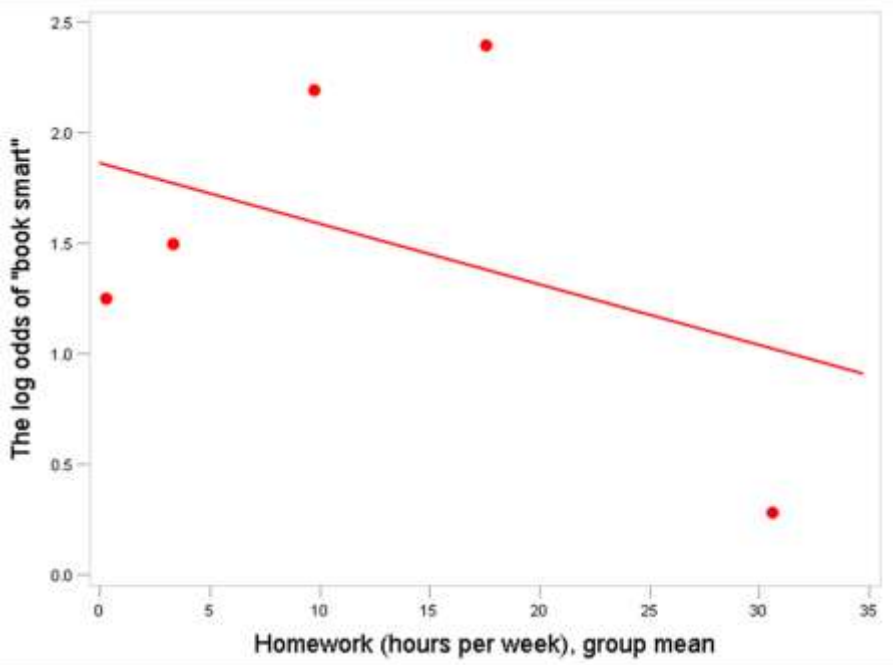


5 groups



What's the approximate equation of the line here?





The logistic regression equation

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	2.1172	0.6357	11.0929	0.0009
homework	1	-0.0389	0.0346	1.2601	0.2616

$$\ln\left(\frac{p}{1-p}\right) = 2.12 - .039 * \text{homework (hours/week)}$$

The logistic model…

$$\boxed{\ln(p/1-p)} = \alpha + \beta_1 * X$$

Logit function =log odds of the outcome

➔ Odds = $\exp[\ln(p/1-p)] = \exp(\alpha + \beta_1 * X)$

And $\exp^{\beta} = OR$

Prediction: predicted logits

For 0 hours per week:

$$\ln\left(\frac{p}{1-p}\right) = 2.12 - .039 * (0) = 2.12$$

For 10 hours per week:

$$\ln\left(\frac{p}{1-p}\right) = 2.12 - .039 * (10) = 1.73$$

For 50 hours per week:

$$\ln\left(\frac{p}{1-p}\right) = 2.12 - .039 * (50) = .17$$

From logits to odds…

For 0 hours per week:

$$\ln\left(\frac{p}{1-p}\right) = 2.12$$

For 10 hours per week:

$$\ln\left(\frac{p}{1-p}\right) = 1.73$$

For 50 hours per week:

$$\ln\left(\frac{p}{1-p}\right) = .17$$

From odds to predicted probabilities

<u>For 0 hours per week</u>:

$$\frac{p}{1-p} = 8.33$$

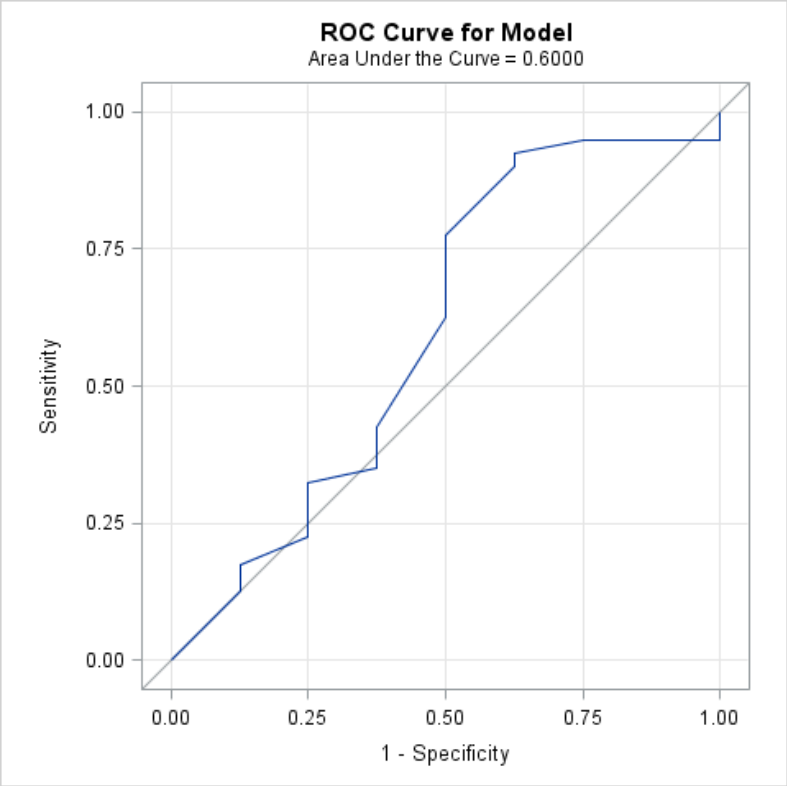
<u>For 10 hours per week:</u>

$$\frac{p}{1-p} = 5.64$$

<u>For 50 hours per week:</u>

$$\ln\left(\frac{p}{1-p}\right) = 1.19$$

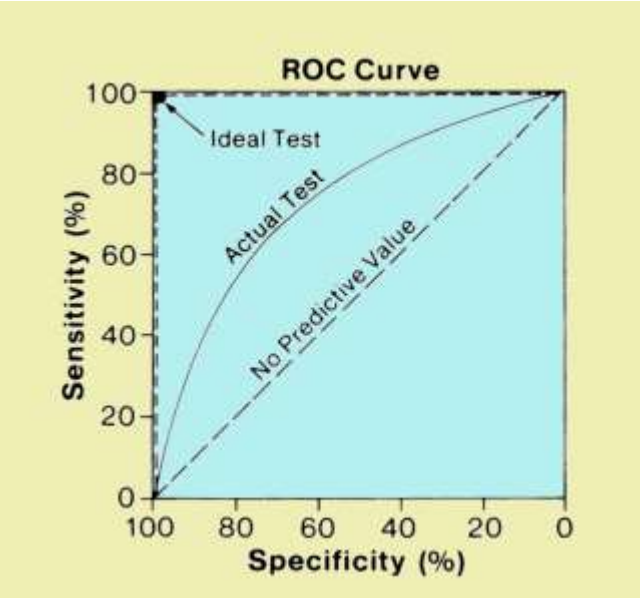
Area under the ROC curve is a measure of model fit…



Area under the ROC curve for homework and book smart example = 60%

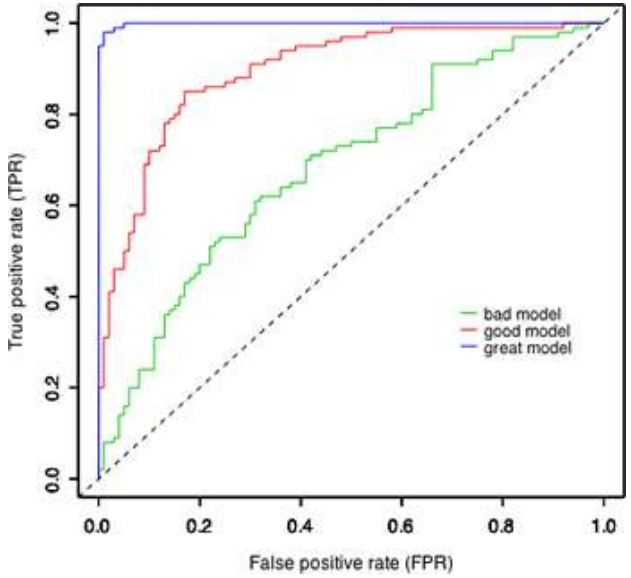
50% means no predictive ability!

ROC (Receiver operating characteristic) curve



FPR = 1- specificity

		Actual Class	
		p	n
Predicted Class	Y	True Positives	False Positives
	N	False Negatives	True Negatives
	Totals:	P	N



What does the beta mean?

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	2.1172	0.6357	11.0929	0.0009
homework	1	-0.0389	0.0346	1.2601	0.2616

using $\exp^{\beta} = OR$

OR = 0.96

From beta to odds ratio…

$$OR = \frac{\text{odds of book smart for higher homework time}}{\text{odds of book smart for lower homework time}}$$

$$= \frac{e^{\alpha + \beta_{\text{hom}\,ework}(1)}}{e^{\alpha + \beta_{\text{hom}\,ework}(0)}} =$$

$$= \frac{\cancel{e^{\alpha}}\, e^{\beta_{\text{hom}\,ework}(1)}}{\cancel{e^{\alpha}}\, e^{\beta_{\text{hom}\,ework}(0)}} = \quad e^{\beta_{\text{hom}\,ework}(1)} = e^{-.039(1)} = 0.96$$

Odds ratio for a continuous predictor…

- Odds ratio of 0.96 for homework means:
- For every 1 hour increase in homework per week, your odds of believing yourself “book smart” decrease by 4% (not significant!).

Multivariate logistic regression

$$\ln(p/1-p) = \alpha + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 \ldots$$

Examples:

ln (odds of lung cancer) = α + β_1*(smoking, yes/no) + β_2*(drinking, yes/no)

ln (odds of lung cancer) = α + β_1*(smoking, yes/no) + β_2*(drinking, yes/no) + β_3*(age)

"Adjusted" Odds Ratio Interpretation (binary predictor)

$$OR = \frac{\text{odds of disease for the exposed}}{\text{odds of disease for the unexposed}}$$

$$= \frac{e^{\alpha + \beta_{alcohol}(1) + \beta_{smoking}(1)}}{e^{\alpha + \beta_{alcohol}(0) + \beta_{smoking}(1)}}$$

$$= \frac{\cancel{e^{\alpha}} e^{\beta_{alcohol}(1)} \cancel{e^{\beta_{smoking}(1)}}}{\cancel{e^{\alpha}} e^{\beta_{alcohol}(0)} \cancel{e^{\beta_{smoking}(1)}}} = \frac{e^{\beta_{alcohol}(1)}}{1} = e^{\beta_{alcohol}(1)}$$

Adjusted odds ratio, continuous predictor

$$OR = \frac{\text{odds of disease for the exposed}}{\text{odds of disease for the unexposed}}$$

$$= \frac{e^{\alpha+\beta_{alcohol}(1)+\beta_{smoking}(1)+\beta_{age}(29)}}{e^{\alpha+\beta_{alcohol}(1)+\beta_{smoking}(1)+\beta_{age}(19)}}$$

$$= \frac{\cancel{e^{\alpha}}\,\cancel{e^{\beta_{alcohol}(1)}}\,\cancel{e^{\beta_{smoking}(1)}}\,e^{\beta_{age}(29)}}{\cancel{e^{\alpha}}\,\cancel{e^{\beta_{alcohol}(1)}}\,\cancel{e^{\beta_{smoking}(1)}}\,e^{\beta_{age}(19)}} = \frac{e^{\beta_{age}(29)}}{e^{\beta_{age}(19)}} = e^{\beta_{age}(10)}$$

Practical Interpretation

$$e^{\beta_{\text{exp}}} = OR_{\text{exposure}}$$

The odds of disease increase *multiplicatively* by $e^{ß}$ for for every one-unit increase in the exposure, controlling for other variables in the model.

Statistics in Medicine

Module 2:

Practice example: Interpreting results from logistic regression

Logistic regression example

- Case-control study of medicine graduates from UCSF.
- Cases= graduates disciplined by the Medical Board of California from 1990-2000 (68).
- Control =graduates (196) were matched by medical school graduation year and specialty choice.
- Aim: "To determine if medical students who demonstrate unprofessional behavior in medical school are more likely to have subsequent state board disciplinary action."

Binary or categorical outcomes (proportions)

Outcome Variable	**Are the observations correlated?**		**Alternatives if sparse data:**
	independent	correlated	
Binary or categorical (e.g. fracture, yes/no)	**Risk difference/relative risks** (2x2 table) **Chi-square test** (RxC table) **Logistic regression** (multivariate regression technique)	**McNemar's chi-square test** (2x2 table) **Conditional logistic regression** (multivariate regression technique) **GEE modeling** (multivariate regression technique)	**McNemar's exact test** (alternative to McNemar's chi-square, for sparse data) **Fisher's exact test** (alternative to the chi-square, for sparse data)

Logistic regression results…

Table 5 Logistic Regression Analysis of Factors Used to Differentiate between 260 Disciplined and Nondisciplined Physician-Graduates of the University of California, San Francisco, School of Medicine, 1990-2000

Table 5

Logistic Regression Analysis of Factors Used to Differentiate between 260 Disciplined and Nondisciplined Physician–Graduates of the University of California, San Francisco, School of Medicine, 1990–2000*

Predictor	Odds Ratio	Confidence Interval (95%)	*p* Value
Men	1.51	0.65–3.51	.34
Undergraduate GPA	.57	0.25–1.28	.17
MCAT lowest quartile	1.01	0.50–2.05	.98
Did not pass ≥ 1 medical school course	1.30	0.59–2.87	.52
Professionalism severity ranking of Concern, Problem, or Extreme	2.15	1.15–4.02	.02

*Predictor variables were coded as follows: male = 0, female = 1; did not pass ≥ 1 course = 0, did pass all courses = 1; MCAT lowest quartile = 0, MCAT not lowest quartile = 1; professionalism rank Concern/Problem/Extreme = 0, Trace/Good = 1. Undergraduate GPA was entered as a continuous variable from 0–4.0.

Statistics in Medicine

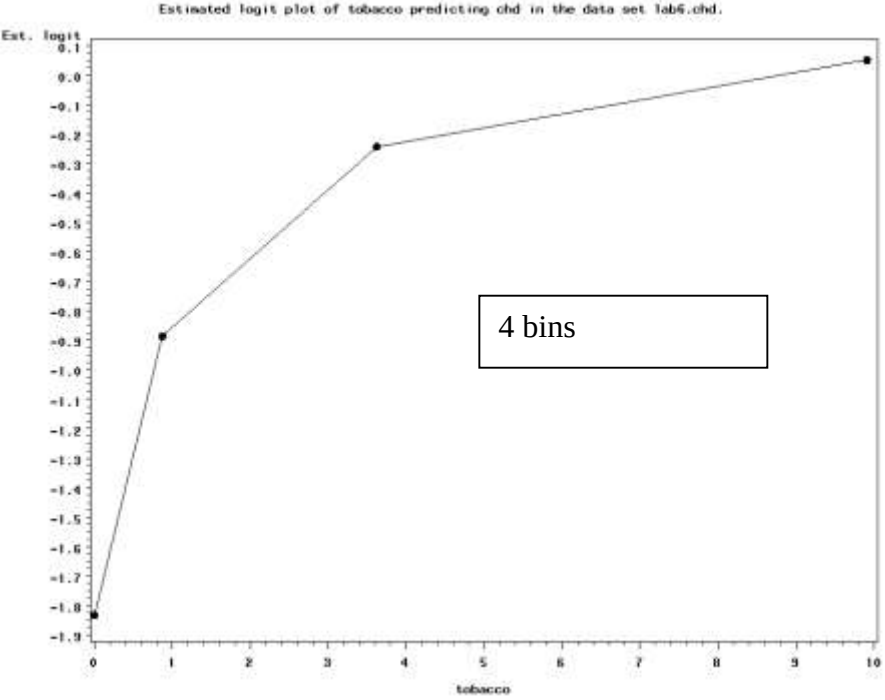
Module 3:

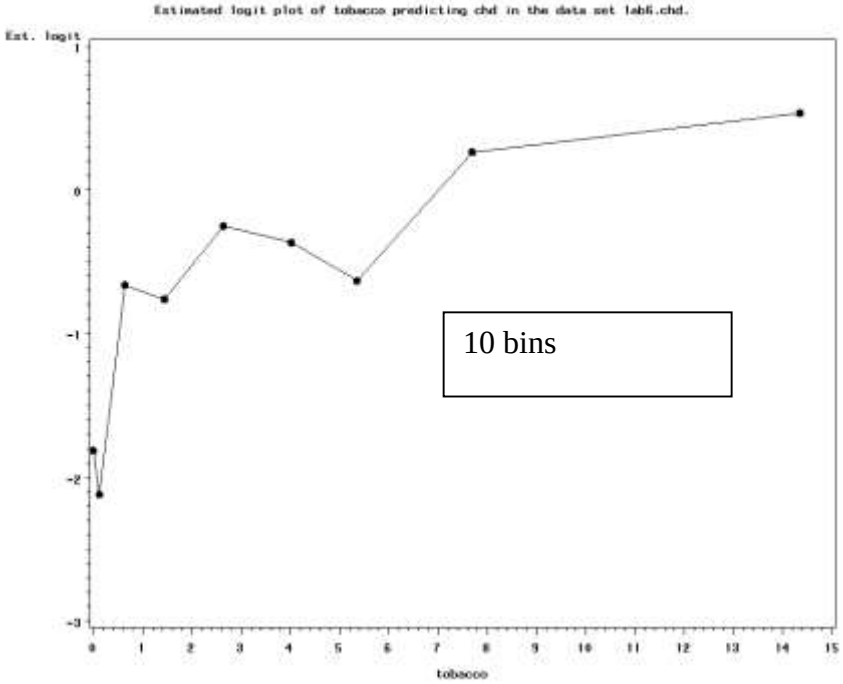
Testing the “linear in the logit” assumption of logistic regression

Linear in the logit

Not linear in the logit… (for continuous data)

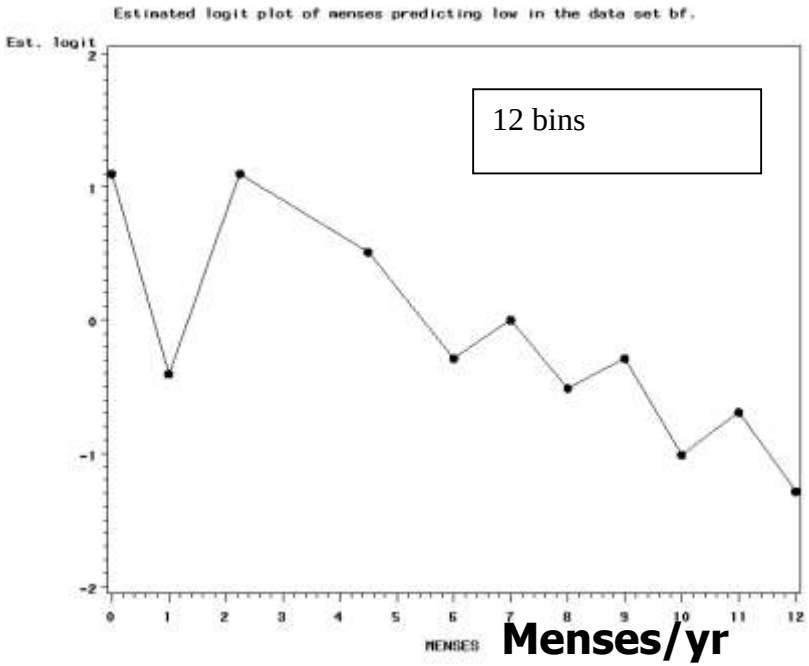
Heart disease vs smoking



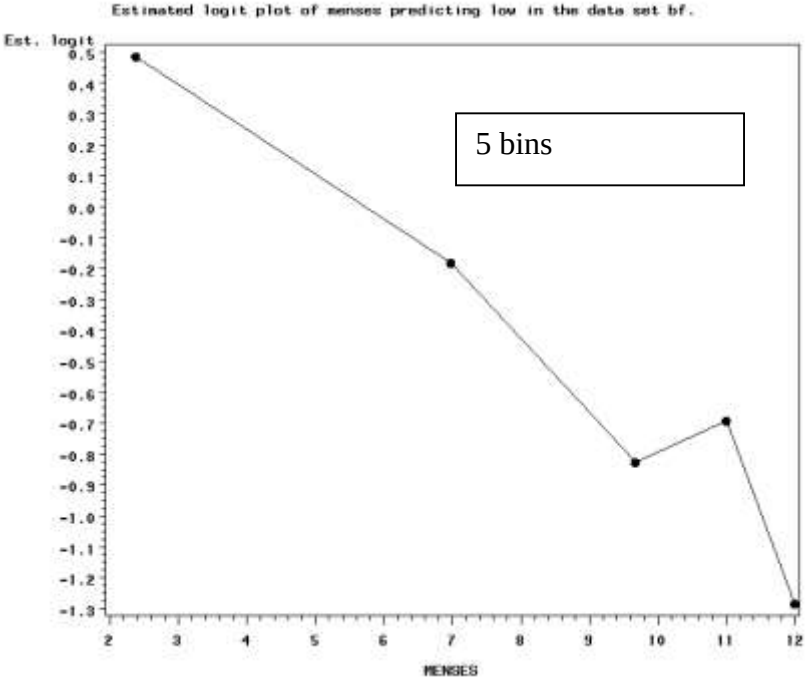


Reasonably linear in the logit…

Bone density of women atheletes



Menses/yr



Statistics in Medicine

Module 4:

Interactions

What is interaction?

- When the effect size (e.g., relationship between a treatment and outcome) is significantly different in different subgroups.
- Example: a blood pressure treatment works significantly better in men than in women(ref).

How do we test for interaction in regression?

- We add an interaction term. If the beta for interaction is significant, this indicates a significant interaction.

Example:

- Blood pressure = $\alpha + \beta_{\text{treatment}}$*(1=drug) + β_{gender}(1=male) + $\beta_{\text{gender*treatment}}$(1 if male and drug)

If $\beta_{\text{gender*treatment}}$ significant ➔ there exits a interaction

Example (interpretation)

1: male
0: female (ref)

1: treatment
0: placebo (ref)

ID	Gender (1,0)	Treatment(1,0)	Interaction (Gender x treatment)
1	1	0	0
2	0	1	0
3	1	1	1

<u>Only men who get treatment has an value "1" for interaction</u>

β_{gender} = effect of gender (male vs female) in blood pressure regardless of treatment

$\beta_{treatment}$ = treatment effect in women (reference) [β_g(0=female), β_{g*t}(0)]

Intercept = mean value of women in baseline [β_g(0=female), β_t*(0=drug), β_{g*t}(0)]

Treatment effect in men is given by $\beta_{treatment} + \beta_{gender*treatment}$

➔ $\beta_{gender*treatment}$ is the difference in treatment effect between men and women (<u>interaction</u>)

Recall: Smoking cessation trial

- Weight-concerned women smokers were randomly assigned to one of four groups:
  - Weight-focused or standard counseling plus bupropion or placebo
- Outcome: biochemically confirmed smoking abstinence

Levine MD, Perkins KS, Kalarchian MA, et al. Bupropion and Cognitive Behavioral Therapy for Weight-Concerned Women Smokers. *Arch Intern Med* 2010;170:543-550.

The Results…

Rates of biochemically verified prolonged abstinence at 3, 6, and 12 months from a four-arm randomized trial of smoking cessation

Months after quit target date	Weight-focused counseling			Standard counseling group		
	Bupropion group (n=106)	Placebo group (n=87)	P-value, bupropion vs. placebo	Bupropion group (n=89)	Placebo group (n=67)	P-value, bupropion vs. placebo
3	41%	18%	.001	33%	19%	.07
6	34%	11%	.001	21%	10%	.08
12	24%	8%	.006	19%	7%	.05

Data excerpted from Tables 2 and 3 of Levine MD, Perkins KS, Kalarchian MA, et al. Bupropion and cognitive behavioral therapy for weight-concerned women smokers. *Arch Intern Med* 2010;170:543-550.

The Results…

Rates of biochemically verified prolonged abstinence at 3, 6, and 12 months from a four-arm randomized trial of smoking cessation

Months after quit target date	Weight-focused counseling			Standard counseling group		
	Bupropion group (n=106)	Placebo group (n=87)	P-value, bupropion vs. placebo	Bupropion group (n=89)	Placebo group (n=67)	P-value, bupropion vs. placebo
3	41%	18%	.001	33%	19%	.07
6	34%	11%	.001	21%	10%	.08
12	24%	8%	.006	19%	7%	.05

Counseling methods appear equally effective in the placebo groups. This implies that there is no main effect for counseling.

The Results…

Rates of biochemically verified prolonged abstinence at 3, 6, and 12 months from a four-arm randomized trial of smoking cessation

Months after quit target date	Weight-focused counseling			Standard counseling group		
	Bupropion group (n=106)	Placebo group (n=87)	P-value, bupropion vs. placebo	Bupropion group (n=89)	Placebo group (n=67)	P-value, bupropion vs. placebo
3	41%	18%	.001	33%	19%	.07
6	34%	11%	.001	21%	10%	.08
12	24%	8%	.006	19%	7%	.05

Bupropion appears to improve quitting rates across both groups. This implies that there is a main effect for drug.

Authors' conclusions/Media coverage…

- "Among weight-concerned women smokers, bupropion therapy increased cessation rates when added to a specialized weight concerns intervention, but not when added to standard counseling"
- The implication: There is an interaction between drug and type of counseling.
- Is there? (Is the effect size much greater in "weight-focused counseling vs the other type of counseling?)

Logistic regression:

Ln (odds of quitting) = α + β_{drug}*(1=drug) + $\beta_{\text{counseling type}}$(1=weight-focused) + $\beta_{\text{drug*counseling type}}$(1 if drug and weight-focused)

Formal test for interaction:

Months after quit target date	Weight-focused counseling		Standard counseling group		**P-value for interaction between bupropion and counseling type**
	Bupropion group (n=106)	Placebo group (n=87)	Bupropion group (n=89)	Placebo group (n=67)	
3	41%	18%	33%	19%	**.42**
6	34%	11%	21%	10%	**.39**
12	24%	8%	19%	7%	**.79**

$\beta_{\text{counseling type}}$, $\beta_{\text{drug*counseling type}}$ both not significant

Sainani KL. Misleading comparisons: the fallacy of comparing statistical significance. *PM&R* 2010; 2 (3): 209-13.

Correct take-home message…

- Bupropion improves quitting rates over counseling alone.
  - Main effect for drug is significant.
  - Main effect for counseling type is NOT significant.
  - Interaction between drug and counseling type is NOT significant.

Example 2

- Cross-sectional study of 1,741 men and women
- Examined relationships between sleep duration, sleep problems, and hypertension (binary outcome).

Vgontzas AN, Liao D, Bixler EO, Chrousos GP, Vela-Bueno A. Insomnia with objective short sleep duration is associated with a high risk for hypertension. *Sleep* 2009;32:491-7.

Example 2: results

	Sleep difficulty	Sleep duration	Sample size	Adjusted OR	95% CI	
reference	Normal sleeping	> 6 h	527	1.00	Low	Upper
	Poor sleep	> 6 h	249	0.79	0.52	1.20
	Insomnia	> 6 h	86	1.31	0.70	2.46
	Normal sleeping	5-6 h	235	0.86	0.60	1.22
	Poor sleep	5-6 h	146	1.48	0.90	2.42
	Insomnia	5-6 h	49	3.53	1.57	7.91
	Normal sleeping	< 5 h	260	1.13	0.79	1.62
	Poor sleep	< 5 h	125	2.43	1.36	4.33
	Insomnia	< 5 h	64	5.12	2.22	11.79

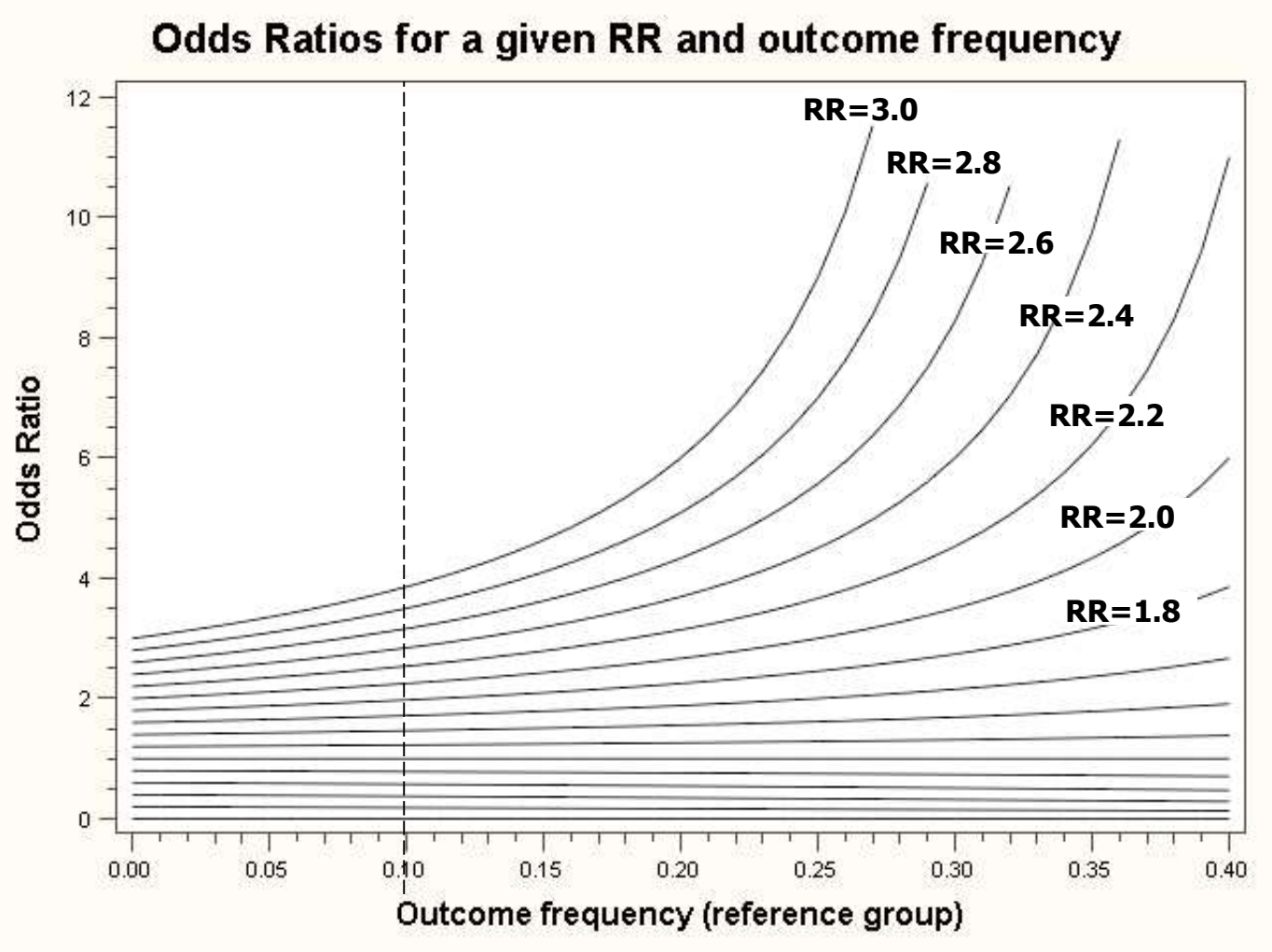
All data adjusted for age, race, sex, BMI, diabetes, smoking status, alcohol consumption, depression, SDB, and sampling weight.

The interaction between insomnia and objective sleep duration is statistically significant, P < 0.01.

Compared to the common reference group, persons without insomnia/ poor sleep and slept more than 6 hours.

model

Don't forget: Odds ratios distort the effect size when the outcome is common!!



Statistics in Medicine

Module 5:

Introduction to Cox regression

Introduction to Cox regression

Outcome Variable	**Are the observation groups independent or correlated?**		**Modifications if assumptions violated:**
	independent	correlated	
Time-to-event (e.g., time to fracture)	**Rate ratio** (2 groups) **Kaplan-Meier statistics** (2 or more groups) **Cox regression** (multivariate regression technique)	**Frailty model** (multivariate regression technique)	**Time-varying effects**

Introduction to Cox Regression

- Also called proportional hazards regression
- Multivariate regression technique where time-to-event (taking into account censoring) is the dependent variable.
- Estimates adjusted hazard ratios.
  - A hazard ratio is a ratio of rates (hazard rates)

Hazard ratios

- A hazard ratio is similar to a rate ratio, but is the ratio of instantaneous incidence rates.
- Since hazard ratios come from a regression, they are usually multivariable-adjusted.

Recall: Ranolazine vs. Placebo

Table 2. Efficacy Outcomes*

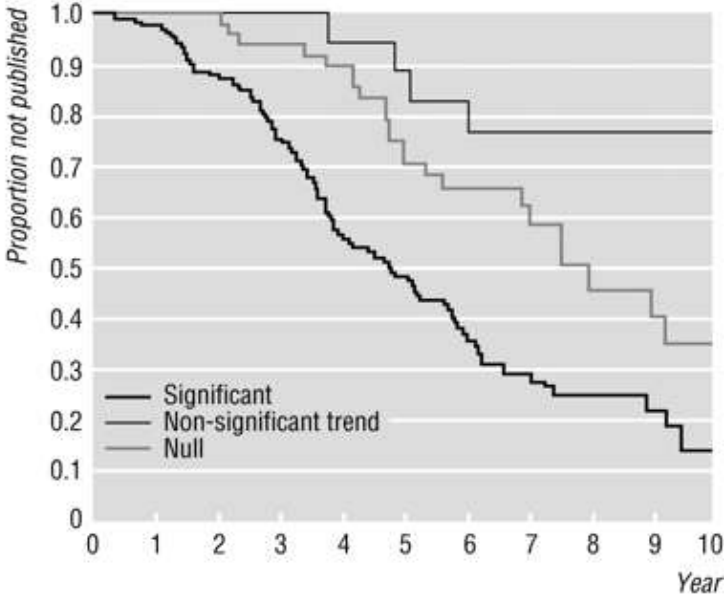
	No. (%) of Patients			
	Ranolazine (n = 3279)	Placebo (n = 3281)	Risk (95% CI)	*P* Value
Randomization to end of study			Hazard Ratio	
Primary end point†	696 (21.8)	753 (23.5)	0.92 (0.83-1.02)	.11
Major secondary end point‡	602 (18.7)	625 (19.2)	0.96 (0.86-1.08)	.50
Cardiovascular death	147 (4.4)	148 (4.5)	1.00 (0.79-1.25)	.98
MI	235 (7.4)	242 (7.6)	0.97 (0.81-1.16)	.76
Recurrent ischemia	430 (13.9)	494 (16.1)	0.87 (0.76-0.99)	.03

Interpretation: the rate of death, MI, or recurrent ischemia (primary end point) was reduced 8% in the ranolazine group compared with placebo (not significant).

Reproduced with permission from: Morrow et al. Effects of Ranolazine on Recurrent Cardiovascular Events in Patients with Non-ST-Elevation Acute Coronary Syndromes. JAMA 2007; 297: 1775-1783.

Example: Study of publication bias

Kaplan-Meier Curve:



No at risk						
Significant	144	127	77	36	15	2
Non-significant trend	20	20	19	14	4	3
Null	52	52	46	24	10	7

Reproduced with permission from: Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects BMJ 1997;315:640-645 (13 September)

Corresponding Cox regression

Table 4 Risk factors for time to publication using univariate Cox regression analysis

Characteristic	# not published	# published	Hazard ratio (95% CI)
Null	29	23	1.00
Non-significant trend	16	4	0.39 (0.13 to 1.12)
Significant	47	99	2.32 (1.47 to 3.66)

Reprduced with permission from: Stern JM, Simes RJ. Publication bias: evidence of delayed publication in a cohort study of clinical research projects BMJ 1997;315:640-645 (13 September)

Interpretation: Significant results have a 2-fold higher incidence of publication compared to null results.

Example 2: Study of mortality in academy award winners for screenwriting

Kaplan-Meier methods

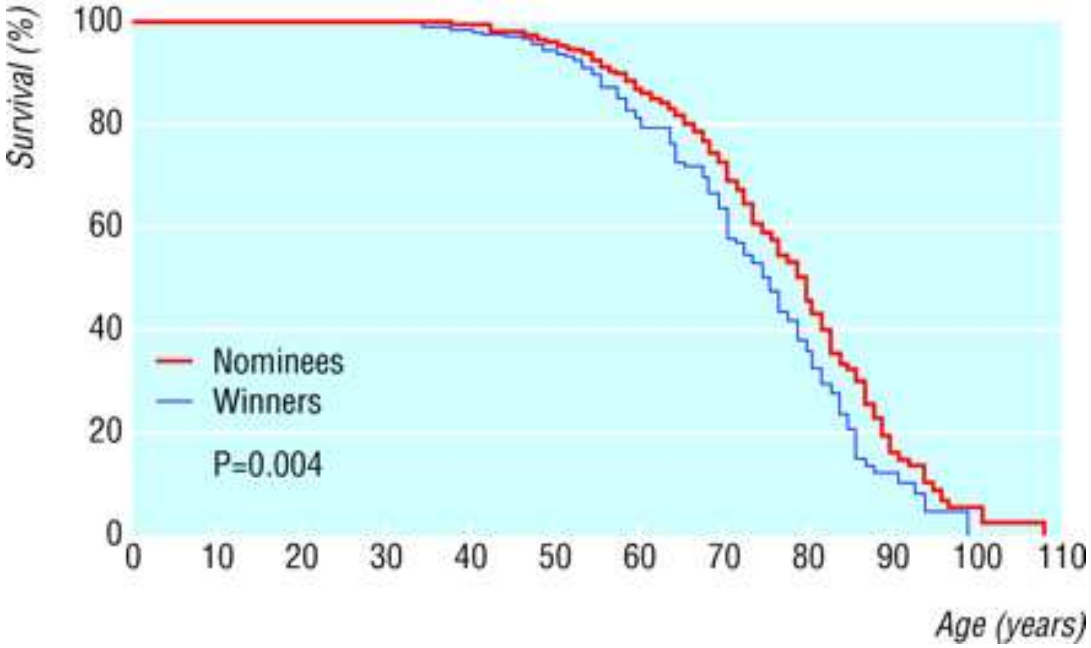


Figure 1 and Table 2 (next slide) were reproduced with permission from: Redelmeier DA, Singh SM. Longevity of screenwriters who win an academy award: longitudinal study. *BMJ* 2001;323:1491-1496 (22-29 December)

Table 2. Death rates for screenwriters who have won an academy award.* Values are hazard ratios (95% confidence intervals) and are adjusted for the factor indicated

	Relative increase in death rate for winners
Basic analysis	1.37 (1.10 to 1.70)
Adjusted analysis	
Demographic:	
Year of birth	1.32 (1.06 to 1.64)
Sex	1.36 (1.10 to 1.69)
Documented education	1.39 (1.12 to 1.73)
All three factors	1.33 (1.07 to 1.65)
Professional:	
Film genre	1.37 (1.10 to 1.70)
Total films	1.39 (1.12 to 1.73)
Total four star films	1.40 (1.13 to 1.75)
Total nominations	1.43 (1.14 to 1.79)
Age at first film	1.36 (1.09 to 1.68)
Age at first nomination	1.32 (1.06 to 1.64)
All six factors	1.40 (1.11 to 1.76)
All nine factors	1.35 (1.07 to 1.70)

HR=1.37; interpretation: 37% higher incidence of death for winners compared with nominees

HR=1.35; interpretation: 35% higher incidence of death for winners compared with nominees even after adjusting for potential confounders

Cox Regression: model details

Linear regression

Logistic regression

Cox regression

<u>Hazard rate</u>

Assumption: constant difference

Cox Regression

- h(t): hazard rate
- h(t): [0 ,1]
- However, intercept of model can go to less than 0, or more than 1

- Therefore, use log
- Ln(h(t)) [-inf, +inf]

The Hazard function

$$h(t) = \lim_{\Delta t \to 0} \frac{R(t) - R(t + \Delta t)}{\Delta t \cdot R(t)}.$$

where R(t) is the survival function

<u>In words:</u> the probability that ***if you survive to t***, you will succumb to the event in the next instant.

The model

Components:

- A baseline hazard function <u>that is left unspecified</u> but must be positive (=the hazard when all covariates are 0)
- A linear function of a set of k fixed covariates

Intercept:
Can take on any form! (not estimated)

$$\ln h_i(t) = \boxed{\ln h_0(t)} + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

Hazard ratio for a binary predictor

e.g. model lung cancer: $\ln h(t) = \ln h_0(t) + \beta_{smoking} + \beta_{age} \quad \Rightarrow e^{\ln h(t)} = h_0(t) e^{\beta_{smoking} + \beta_{age}}$

smoker

$$HR_{lung\ cancer/smoking} = \frac{h_i(t)}{h_j(t)} = \frac{\cancel{h_0(t)} e^{\beta_{smoking}(1) + \cancel{\beta_{age}(60)}}}{\cancel{h_0(t)} e^{\beta_{smoking}(0) + \cancel{\beta_{age}(60)}}} = e^{\beta_{smoking}(1-0)}$$

Non-smoker

$$HR_{lung\ cancer/smoking} = e^{\beta_{smoking}}$$

This is the hazard ratio for smoking adjusted for age.

Hazard ratio for a continuous predictor

$$HR_{lung\ cancer/10-years\ \text{increase in age}} = \frac{h_i(t)}{h_j(t)} = \frac{\cancel{h_0(t)}\cancel{e^{\beta_{smoking}(0)}}e^{\beta_{age}(70)}}{\cancel{h_0(t)}\cancel{e^{\beta_{smoking}(0)}}e^{\beta_{age}(60)}} = e^{\beta_{age}(70-60)}$$

$$HR_{lung\ cancer/10-years\ \text{increase in age}} = e^{\beta_{age}(10)}$$

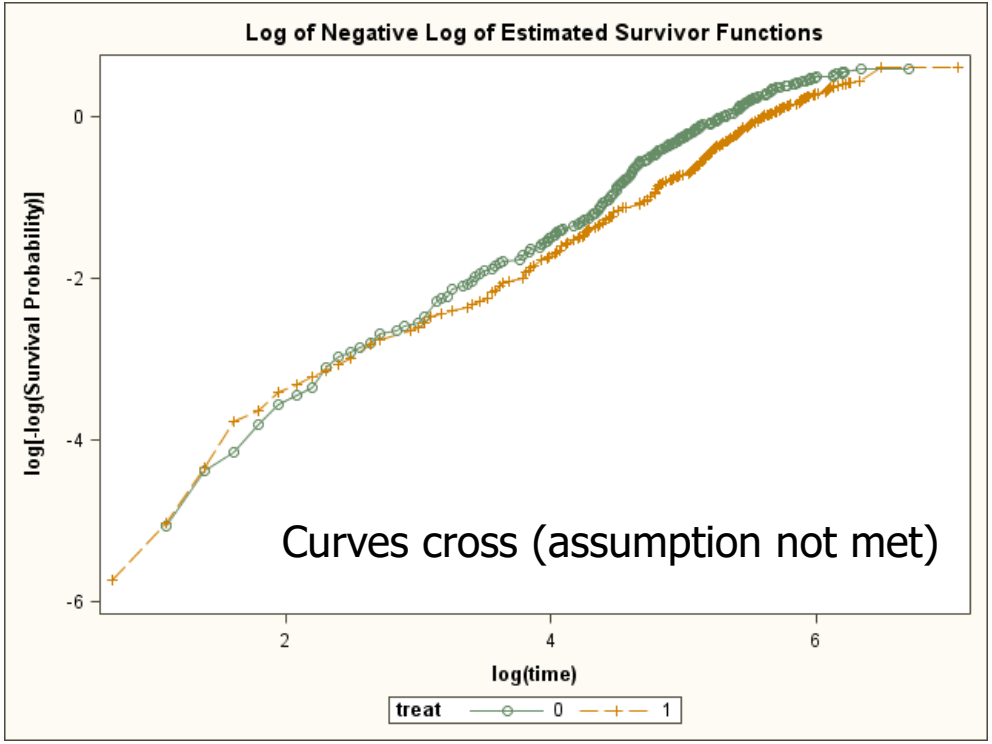
This is the hazard ratio for a 10-year increase in age, adjusted for smoking.

Exponentiating a continuous predictor gives you the hazard ratio for a 1-unit increase in the predictor.

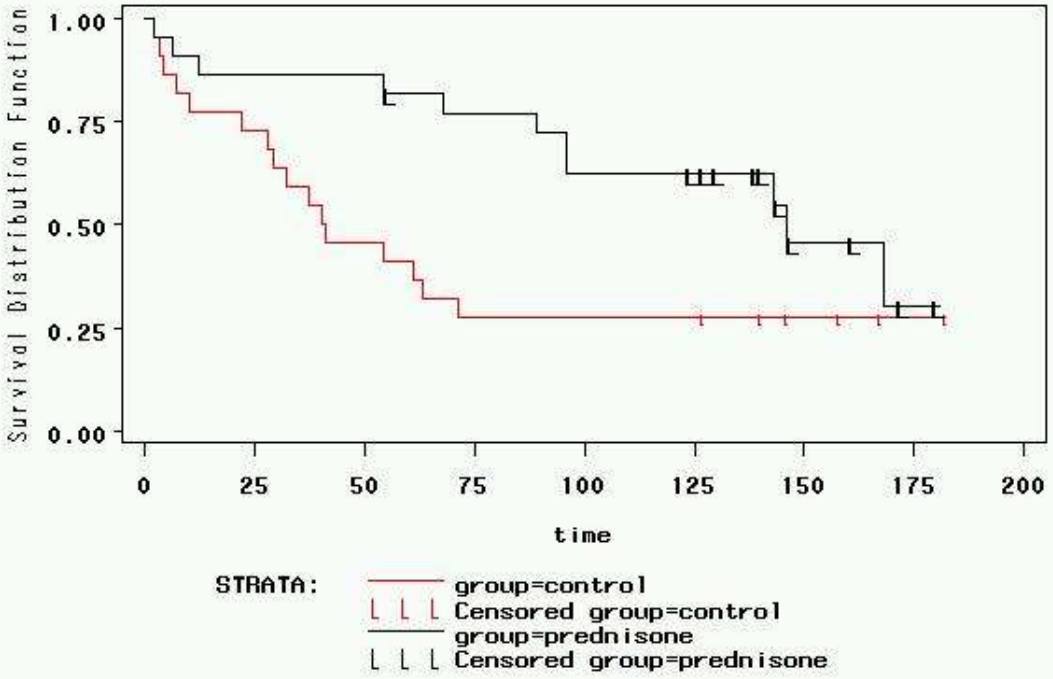
The Proportional Hazards Assumption

Output is single HR

Testing Proportional hazards: e.g. log-log plot



Recall: Hepatitis Example



Data reproduced with permission from: Bland and Altman. Time to event (survival) data. *BMJ* 1998;317:468.

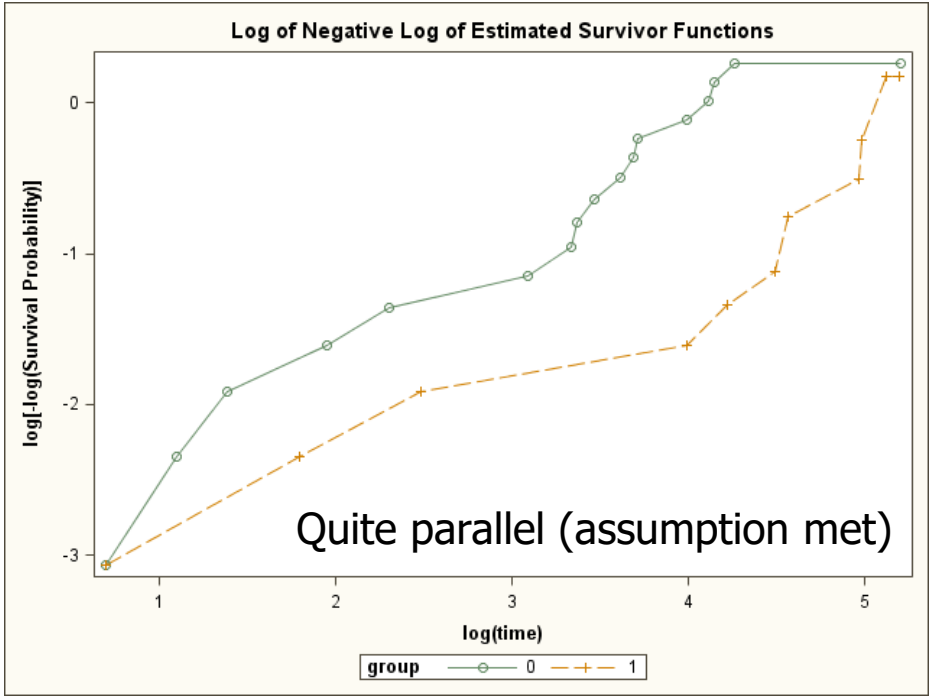
Corresponding Cox regression

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Treatment vs. Control	1	-0.83230	0.39739	4.3865	0.0362	0.435

Note: No intercept

Meaning of HR: 57% decrease in mortality rate for patients on drug

Test of proportional hazards assumption: log-log plot



Other tests available

Statistics in Medicine

Module 6:

Regression worries: Residual confounding

Residual confounding

- You cannot completely wipe out confounding simply by adjusting for variables in multiple regression unless variables are measured with zero error (which is usually impossible).
- Example: meat eating and mortality

Men who eat a lot of meat are unhealthier for many reasons!

Table 1. Selected Age-Adjusted Characteristics of the National Institutes of Health–AARP Cohort by Red Meat Quintile Category[a]

Characteristic	Red Meat Intake Quintile, g/1000 kcal				
	Q1	Q2	Q3	Q4	Q5
Men (n=322 263)					
Meat intake					
Red meat, g/1000 kcal	9.3	21.4	31.5	43.1	68.1
White meat, g/1000 kcal	36.6	32.2	30.7	30.4	30.9
Processed meat, g/1000 kcal	5.1	7.8	10.3	13.3	19.4
Age, y	62.8	62.8	62.5	62.3	61.7
Race, %					
Non-Hispanic white	88.6	91.8	93.1	94.0	94.1
Non-Hispanic black	4.2	3.2	2.7	2.2	1.9
Hispanic/Asian/Pacific Islander/American Indian/Alaskan native/unknown	7.2	5.0	4.2	3.8	4.0
Positive family history of cancer,%	47.0	47.7	48.4	48.6	47.8
Currently married, %	80.8	84.4	86.1	86.7	85.6
BMI	25.9	26.7	27.1	27.6	28.3
Smoking history, %[b]					
Never smoker	34.4	30.5	28.8	27.6	25.4
Former smoker	56.5	58.1	57.5	57.1	55.8
Current smoker or having quit <1 y prior	4.9	7.6	9.9	11.4	14.8
Education, college graduate or postgraduate, %	53.0	47.3	45.1	42.3	39.1
Vigorous physical activity ≥5 times/wk, %	30.7	23.6	20.5	18.6	16.3
Dietary intake					
Energy, kcal/d	1899	1955	1998	2038	2116
Fruit, servings/1000 kcal	2.3	1.8	1.6	1.4	1.1
Vegetables, servings/1000 kcal	2.4	2.1	2.0	2.0	1.9

Reproduced with permission from: Sinha R, Cross AJ, Graubard BI, Leitzmann MF, Schatzkin A. Meat intake and mortality: a prospective study of over half a million people. *Arch Intern Med* 2009;169:562-71

Mortality risks…

Table 2. Multivariate Analysis for Red, White, and Processed Meat Intake and Total and Cause-Specific Mortality in Men in the National Institutes of Health–AARP Diet and Health Study[a]

Mortality in Men (n=322 263)	Quintile Q1	Q2	Q3	Q4	Q5	*P* Value for Trend
	Red Meat Intake[b]					
All mortality						
Deaths	6437	7835	9366	10 988	13 350	
Basic model[c]	1 [Reference]	1.07 (1.03-1.10)	1.17 (1.13-1.21)	1.27 (1.23-1.31)	1.48 (1.43-1.52)	<.001
Adjusted model[d]	1 [Reference]	1.06 (1.03-1.10)	1.14 (1.10-1.18)	1.21 (1.17-1.25)	1.31 (1.27-1.35)	<.001
Cancer mortality						
Deaths	2136	2701	3309	3839	4448	
Basic model[c]	1 [Reference]	1.10 (1.04-1.17)	1.23 (1.16-1.29)	1.31 (1.24-1.39)	1.44 (1.37-1.52)	<.001
Adjusted model[d]	1 [Reference]	1.05 (0.99-1.11)	1.13 (1.07-1.20)	1.18 (1.12-1.25)	1.22 (1.16-1.29)	<.001
CVD mortality						
Deaths	1997	2304	2703	3256	3961	
Basic model[c]	1 [Reference]	1.02 (0.96-1.08)	1.10 (1.04-1.17)	1.24 (1.17-1.31)	1.44 (1.37-1.52)	<.001
Adjusted model[d]	1 [Reference]	0.99 (0.96-1.09)	1.08 (1.02-1.15)	1.18 (1.12-1.26)	1.27 (1.20-1.35)	<.001
Mortality from injuries and sudden deaths						
Deaths	184	216	228	280	343	
Basic model[c]	1 [Reference]	1.02 (0.84-1.24)	0.97 (0.80-1.18)	1.09 (0.90-1.31)	1.24 (1.03-1.49)	.01
Adjusted model[d]	1 [Reference]	1.06 (0.86-1.29)	1.01 (0.83-1.24)	1.14 (0.94-1.39)	1.26 (1.04-1.54)	.008
All other deaths						
Deaths	1268	1636	1971	2239	2962	
Basic model[c]	1 [Reference]	1.13 (1.05-1.22)	1.25 (1.17-1.35)	1.33 (1.24-1.42)	1.68 (1.57-1.80)	<.001
Adjusted model[d]	1 [Reference]	1.17 (1.09-1.26)	1.28 (1.19-1.38)	1.34 (1.25-1.44)	1.58 (1.47-1.70)	<.001

Unadjusted (significant)
Adjusted (significant)

Cancer only (significant)

??? Likely due to residual confounding

Reproduced with permission from: Sinha R, Cross AJ, Graubard BI, Leitzmann MF, Schatzkin A. Meat intake and mortality: a prospective study of over half a million people. *Arch Intern Med* 2009;169:562-71

Residual confounding

- For a binary predictor, incomplete of confounding can plausibly generate spurious relative risks in the range of 0.6 to 1.6.
- In addition to creating spurious associations, residual confounding can also obscure relationships, leading researchers to miss associations.