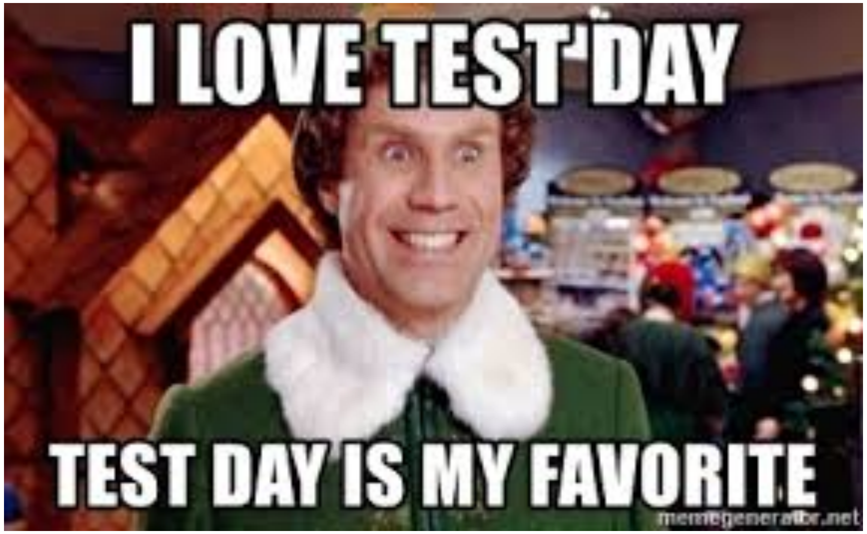
BST 210

Applied Regression Analysis



Final Exam

- **Tuesday, December 17**
- **11:30am – 1:00pm**
- FXB G12 and FXB G13
- Open book, note, computer (limit to Canvas if online)

  NOTE: while this exam is 'open-everything', it is suggested that you review extensively and create a summary sheet (perhaps front and back) for quicker reference
- Be able to exponentiate or take logarithm

Lecture 28
Plan for Today

Course review!

**** Note: be on lookout for quick survey coming today**

BST 210 Course Overview

This course covered the major areas of regression analysis, including methods of analysis for data of these forms:

- Continuous
- Categorical
- Count
- Time to event (survival outcomes)

Such methods of regression analysis include:

- Linear regression
- Logistic regression and extensions
- Poisson regression and extensions
- Survival curve estimation, log rank tests, and Cox proportional hazards regression

Regression	Outcome	Assumptions	Model	Effect Estimate
Linear	Continuous	Linearity, Independence, Normality, & Equal variance	$E[Y_i \| X_i] = \beta_0 + \beta_1 \cdot X_i$	β_1 is the change in $E[Y_i]$ associated with a one unit change in X
Logistic	Binary		$\text{logit}(P(Y_i = 1)) = \beta_0 + \beta_1 X_i$	e^{β_1} is the odds ratio associated with a one unit change in X
Multinomial	Categorical		$\log(\frac{P(Y=k)}{P(Y=0)}) = \beta_{k0} + \beta_{k1} X_i$	$e^{\beta_{k1}}$ is the relative risk ratio of being in outcome group k as compared to outcome group 0 for a one unit change in X
Ordinal	Ordinal	Proportional odds	$\log\left(\frac{P(Y \geq j)}{P(Y < j)}\right) = \beta_{0j} + \beta_1 X_i$	e^{β_1} is the odds ratio for $Y \geq j$ versus $Y < j$ associated with a one unit change in X. Note that this is the same for all cut-points j.
Poisson	Count Data	$E[Y_i] = Var(Y_i)$, Incidence rate λ is time invariate	$\log(E[Y_i \| X_i]) = \beta_0 + \beta_1 X_i$ $+\log(t_i)$	e^{β_1} is the incidence rate ratio associated with a one unit increase in X
Survival (Cox)	Time-to-Event	Proportional hazards Also, Exponential: constant baseline hazard Weibull: Weibull baseline hazard	$\lambda(t \| X) = \lambda_0(t)\exp\{\beta_0 + \beta_1 X_i\}$	e^{β_1} is the hazard ratio associated with a one unit increase in X

BST 210 Course Overview

Many additional themes went through the course;

- Assessment of confounding and effect modification
- Model building and model selection
- Assessment of goodness-of-fit
- Interpretation of models and presentation of results

BST 210 Course Overview

- Refer also to our Framework Tree!

BST 210 Course Overview

We included some modern data analysis methods;

- Generalized additive models and restricted cubic splines, to assess linearity of effect in continuous covariates
- Modern methods for model selection (AIC, BIC, mention of bootstrapping)
- Generalized linear models and overdispersion
- Complex sample size/power calculations
- Censored observations and missing data

BST 210 Course Overview

We explored varied examples, including

- Epidemiologic studies
- Clinical trials
- Real-life examples
- Design of new studies
- *At least* 50 datasets and example studies—some of the most influential studies across many fields where regression modeling is relevant

Course Themes: Regression Methods

- Linear regression for continuous outcomes
- Logistic regression for binary outcomes (or other links)
- Multinomial regression for $k \geq 3$ unordered outcomes
- Ordinal regression for $k \geq 3$ ordered outcomes
- Conditional logistic regression for matched outcomes (or small/sparse strata)
- Poisson regression for count outcomes, and extensions
- Cox proportional hazards, Exponential, and Weibull regression for survival outcomes

Course Themes: Confounding

- Definition of confounding
- Mantel-Haenszel method for stratified 2x2 tables (for odds ratios, but similar methods for risk ratios and risk differences exist, and for stratified incidence rate ratios for count data)
- Log-rank test for stratified survival outcomes
- Regression adjustment – often more flexible

Course Themes: Effect Modification

- Definition of effect modification

- Mantel-Haenszel method for stratified 2x2 tables (for odds ratios, but similar methods for risk ratios or risk differences exist, and for stratified incidence rate ratios for count data)

- Regression methods using interaction terms – often more flexible

Course Themes: Model Selection

- Answer your research question
- Model what others have done in the literature
- Use logic and common sense
- Wald tests can easily compare single parameters
- Likelihood ratio tests can be used to compare nested models
- Model building procedures generally focus on *p*-values and parsimony
- Model building can also use AIC, BIC, or other methods (adjusted R^2, MSE, G-O-F, …)

Course Themes: Model Selection

- Always consider appropriate confounding and effect modification questions

- Assess goodness-of-fit when possible

- More than one final model may be appropriate and presented in an analysis or paper

Things to Know

- When to use each method
  - Advantages, disadvantages, and assumptions
- How to interpret the output
  - All the models: coefficients, CIs, GOFs, etc
- Types of missing data
  - Pros and cons of different approaches for missing data
- Review everything—these are just suggestions!

Review: conditional logistic regression

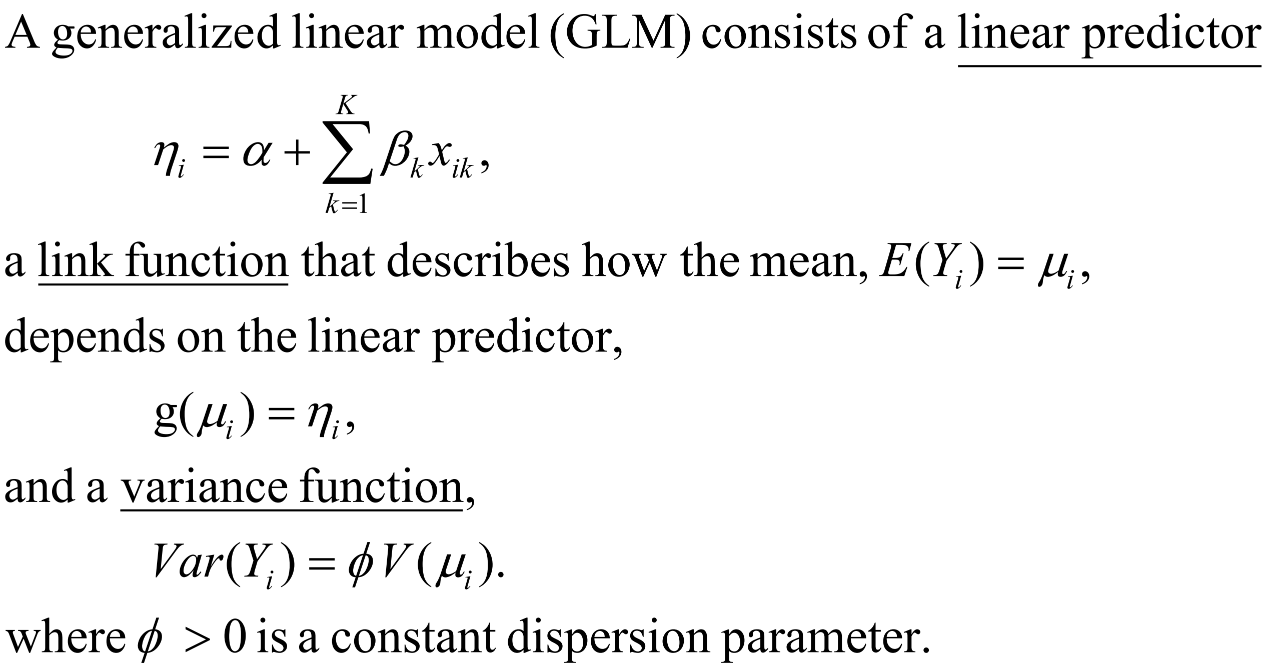
- Since the matched sets are usually small, it is virtually impossible to directly estimate the α_i, since there are so many α_i values

- Use a conditional approach to estimate the other regression parameters (i.e., β_k, $k = 1, \ldots, K$), “conditioning out” the effect of the α_i parameters

- Use conditional maximum likelihood estimates (CMLEs) that maximize this conditional likelihood $L_C(\beta)$

Review: generalized linear models

- Way to view all of the models in a larger context
- What are the 3 ingredients?
- Exponential family? Canonical?



A generalized linear model (GLM) consists of a linear predictor

$$\eta_i = \alpha + \sum_{k=1}^{K} \beta_k x_{ik},$$

a link function that describes how the mean, $E(Y_i) = \mu_i$, depends on the linear predictor,

$$\mathrm{g}(\mu_i) = \eta_i,$$

and a variance function,

$$Var(Y_i) = \phi V(\mu_i).$$

where $\phi > 0$ is a constant dispersion parameter.

Review: Poisson regression

- Poisson regression can be applied when the response variable is a count that follows a Poisson distribution (the analysis of incidence rates)

- The incidence rate can be a function of covariates, but should be constant over time conditional on the values of those covariates (**stationarity**)

- Follow-up times are allowed to vary for individual subjects
- The number of events for each subject is assumed to be independent of other subjects

Review: Poisson Regression Model

Assume Y_i is a Poisson random variable, where

$$E(Y_i) = \mu_i = \lambda_i t_i, \text{ and}$$

$$\log(\lambda_\mathrm{i}) = \alpha + \sum_{k=1}^{K} \beta_k x_{ik}, \text{ where}$$

$$x_{ik} = kth \text{ covariate for the } ith \text{ subject}$$

The incidence rate is a function of one or more covariates

Expressing the model in terms of μ,

$$\begin{aligned} \log(\mu_i) &= \log(\lambda_i) + \log(t_i) \\ &= \alpha + \sum_{k=1}^{K} \beta_k x_{ik} + \log(t_i) \end{aligned}$$

The term $\log(t_\mathrm{i})$ is called an *offset*

Review: Poisson regression

- Goodness of fit: (want to fail to reject)

H_0: the model fits the data (well-calibrated) versus

H_A: the model does not fit the data

  - Pearson chi square statistic
  - Deviance statistic
    - It has the same degrees of freedom as the Pearson test, and is equivalent to the likelihood ratio test comparing your model with the saturated model (containing J parameters)

Usually the Pearson and deviance statistics will generally give you similar findings, though they can only be used for small to moderate number of covariate patterns

Review: Poisson regression

- One (strong) assumption of Poisson regression is that the mean and variance of a Poisson outcome are equal

- Overdispersion:
  - negative binomial regression
  - Robust variance estimator

- More “zero outcomes” than expected:
  - *zero-inflated* Poisson model

Review: Survival Analysis

- In some studies, the response variable of interest is the length of time between an initial observation and the occurrence of a subsequent event (survival time)

- A nonparametric method for estimating survival curves that takes censoring into account is called the *Kaplan-Meier estimator* or the *product-limit estimator*

Review: Survival Function

- A random variable T represents the time from a start point to an event of interest (e.g., time from start of treatment to serious adverse event, time from disease remission to recurrence)

- By definition, T must be ≥ 0

- The *survival function* $S(t)$ is defined as:

  - $S(t) = P(T > t)$

- It is the proportion of individuals who are event-free at time t

Review: Hazard Function

- The *hazard function* (or hazard rate) is defined as the instantaneous rate of failure at time *t*, given that a subject has survived to time *t*

- Mathematically, it can be defined by:

$$h(t) = \lim_{\Delta t \to 0} [\Pr(t \le T \le t + \Delta t) / \Delta t \,|\, T \ge t]$$

- Note that *h*(*t*) ≥ 0 and has no upper bound; it is not a probability

Review: Kaplan-Meier Estimate

$$\hat{S}(t) = 1 \; for \; 0 \le t < t_1,$$

$$\hat{S}(t) = \hat{\Pr}(T > t_1 \,|\, in\ risk\ set\ at\ time\ t_1)$$

$$= 1 - d_1 / n_1 = \hat{q}_1 \; for \; t_1 \le t < t_2,$$

$$\hat{S}(t) = \hat{\Pr}(T > t_1 \,|\, in\ risk\ set\ at\ time\ t_1) \text{ x } \hat{\Pr}(T > t_2 \,|\, in\ risk\ set\ at\ time\ t_2)$$

$$= (1 - d_1 / n_1)(1 - d_2 / n_2) = \prod_{j=1}^{2} (1 - d_j / n_j) \; for \; t_2 \le t < t_3.$$

In general, the survival function is estimated as:

$$\hat{S}(t) = \prod_{j=1}^{k} (1 - d_j / n_j) = \prod_{j=1}^{k} \hat{q}_j \; for \; t_k \le t < t_{k+1}.$$

Review: Log-Rank Test

- To compare two survival curves, the most commonly used test is a generalization of the Mantel-Haenszel test for stratified 2×2 tables known as the *log-rank test*

- The log-rank test is used to test the hypothesis that two survival distributions are equal

- H_0: $S_1(t) = S_2(t)$ for all t

  H_1: $S_1(t) \neq S_2(t)$

- The alternative hypothesis means that the two survival functions differ for at least one value of t, but we cannot be more specific than this

Review: Regression Models for Survival Data

Exponential Survival Model (hazard assumed constant over time):

$$h(t|X_1 = x_1, \ldots, X_p = x_p) = h_0 \exp(\beta_1 x_1 + \ldots + \beta_p x_p)$$

$$h(t|x_1 = 0, \ldots, x_p = 0) = \exp(\beta_0) \equiv h_0(t) = h_0$$

Weibull Survival Model (hazard can vary over time):

$$h(t) = h_0(t) \exp(\beta_1 x_1 + \ldots + \beta_p x_p),$$

where

$$h_0(t) = \gamma t^{\gamma-1} \lambda. \qquad h_0(t) = \exp(\beta_0) \gamma t^{\gamma-1}$$

Cox Proportional Hazard (don’t specify h_0):

$$h(t|X_1 = x_1, \ldots, X_p = x_p) = h_0 \exp(\beta_1 x_1 + \ldots + \beta_p x_p)$$

27

Review: Cox Proportional Hazards Model

The general proportional hazards model is:

$$h(t \vert X_1 = x_1, \ldots, X_p = x_p) = h_0(t) \exp(\beta_1 x_1 + \ldots + \beta_p x_p)$$

which we can rewrite in the form:

$$\log[h(t \vert \mathrm{X}_1 = x_1, \ldots, X_p = x_p)] = \log[h_0(t)] + \beta_1 x_1 + \ldots + \beta_p x_p,$$

where $h_0(t)$ is the baseline hazard function and the intercept term is included in $\log[h_0(t)]$.

Review: Cox Proportional Hazards Model

- Here β_i is the log hazard ratio associated with a one unit increase in x_i, holding all the other variables constant

- Also, exp(β_i) is the hazard ratio associated with a one unit increase in x_i, holding all the other variables constant

- An assumption of the Cox model is that the hazards of risk factors are proportional over time

- Specifically, when comparing two groups (e.g., with a binary covariate) and holding all the other covariates constant, the hazard ratio of exposed versus control is assumed to be the same across all times

29

Review: Checking the Proportional Hazards Assumption

- If we plot the complementary log-log transformation of $S(t)$ versus log(t) for each group, the graphs should be roughly parallel and separated by a constant β (Stata plots this with one more multiplication by – 1)
- This is often called a *log-log plot*
- It is used to *visually evaluate* the assumption of proportional hazards
- Also, check KM curves shouldn’t cross

Review: Power and Sample Size

- Power = P(reject H_0 | H_0 false) = 1 – β

  β = P(fail to reject H_0 | H_0 false)= P(Type II error)

  α = P(reject H_0 | H_0 true) = P(Type I error)

  = significance level of test

- Need to make a lot of assumptions
- Good to give a range of values
- Account for drop out or non compliance

Review: Sample Size Estimation

- To detect a hazard ratio of HR, the total number of events required (using Freedman's method) is:

$$d = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 \, (k \text{ HR} + 1)^2}{k \, (\text{HR} - 1)^2}$$

- This is for a two-sided α level test with power $1 - \beta$.

Review: Missing Data

- Missing completely at random (MCAR) if the missingness (R) is independent of both *Y* and *x*
  - $P(R = 1 \mid x, Y) = P(R = 1)$

- Missing at random (MAR) given the covariate *x* if the missingness is independent of *Y* given *x*
  - $P(R = 1 \mid x, Y) = P(R = 1 \mid x)$
  - $f(Y \mid x, R = 0) = f(Y \mid x, R = 1)$
  - Need maximum likelihood based methods

- Missing not at random (MNAR) if they are neither MCAR nor MAR
  - $f(Y \mid x, R = 0) \neq f(Y \mid x, R = 1)$

Solutions: exclude missing data, impute

Practice Example (from last time)

- For each example below for the outcome BMI and the covariate age, describe the type of missingness (MCAR, MAR, MNAR) and how to adjust for this.

1. BMI is missing for some subjects because only one scale was available. Some subjects didn’t have their weight measured if the scale was occupied.

2. BMI is missing for younger subjects because they had to go to school and couldn’t wait to have their BMI measured.

3. BMI is missing for older subjects with a higher BMI because they didn’t want their weight measured.

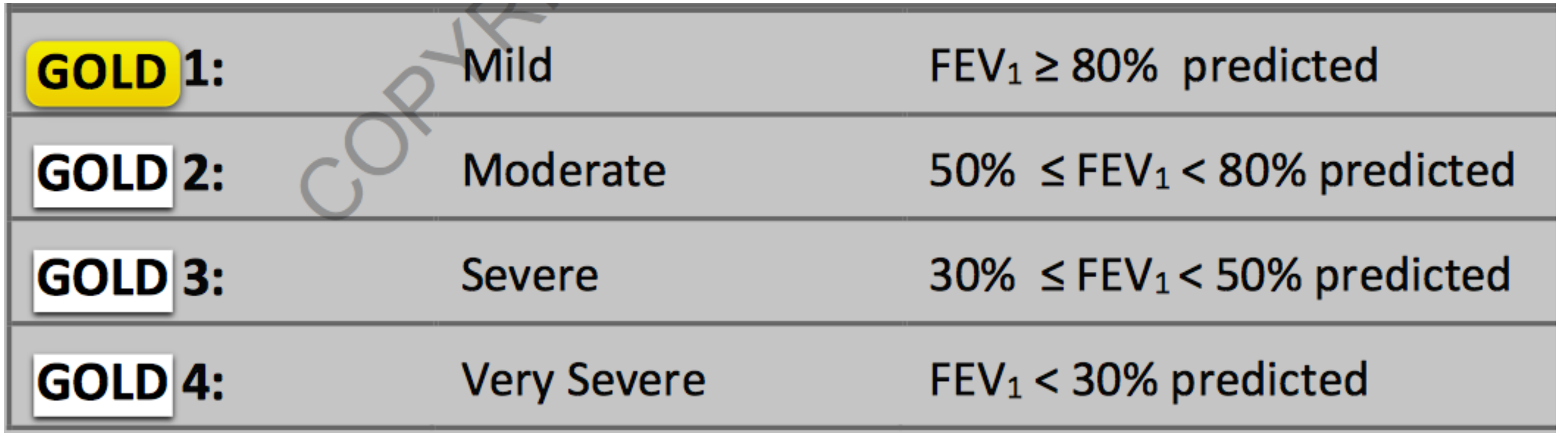
Practice (1) : Which Method to use?

- Is there an association with the continuous outcome FEV$_1$ and current smoking status adjusting for age?
  - FEV$_1$ is the maximal amount of air you can forcefully exhale in one second

- What assumptions or issues should you consider for this method?

Practice (2): Which Method to use?

- For subjects with COPD, is there an association with the outcome Gold Stage and current smoking status adjusting for age?



GOLD 1:	Mild	$FEV_1 \geq 80\%$ predicted
GOLD 2:	Moderate	$50\% \leq FEV_1 < 80\%$ predicted
GOLD 3:	Severe	$30\% \leq FEV_1 < 50\%$ predicted
GOLD 4:	Very Severe	$FEV_1 < 30\%$ predicted

- What assumptions or issues should you consider for this method?

Practice (3) : Which Method to use?

- Is there an association with time to death and current smoking status adjusting for age?

- What assumptions or issues should you consider for this method?

Practice (4): Which Method to use?

- Is there an association with number of hospitalizations for COPD in a year and current smoking status adjusting for age?

- What assumptions or issues should you consider for this method?

Practice Problem (5)

- A recent study looked at the survival time (in months) of 35 subjects with very high cholesterol levels (> 340 mg/100 ml) at entry.

- A proportional hazards regression model was run on the data, yielding the following results:

Covariate	Hazard Ratio	95% Confidence Interval
Cigs/day (continuous)	1.0504	0.9820, 1.1236
I(cholesterol ≥ 400)	4.888	0.913, 26.169

- Using this model, what is the estimate of the hazards ratio associated with a 20 cigarette/day increase in smoking for someone whose cholesterol level is 450? If possible, also calculate a 95% confidence interval for this quantity, or else briefly explain why this is not possible.

Practice Problem (5)

- If possible based on the information available, calculate the two-sided *P*-value for the effect of smoking, adjusting for I(cholesterol ≥ 400). Otherwise, briefly explain why this is not possible.

Practice Problem (5)

- One additional covariate to consider was blood pressure, but unfortunately 10 of the 35 subjects had missing blood pressures.

- Briefly describe some descriptive analyses you might perform that could inform your inferences about adding the effects of blood pressure to your prediction model.

Practice Problem (5)

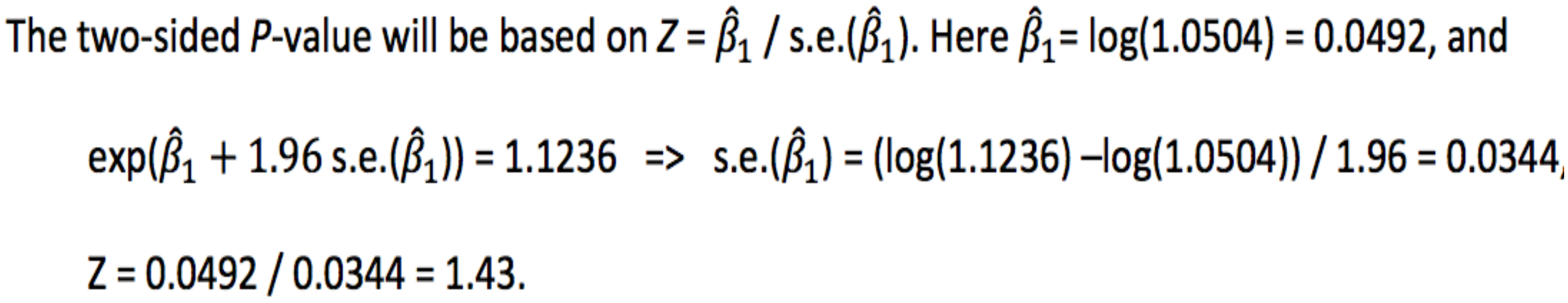
- The hazard ratio estimate for a 20 cigarette/day increase in smoking for someone whose cholesterol is 450 only involves the smoking effect (since there is no interaction with cholesterol in this model)

- estimate of the hazards ratio =

$$\exp(20 \times \hat{\beta}_1) = (\exp(\hat{\beta}_1))^{20} = (1.0504)^{20} = 2.674.$$

- Similarly, an approximate 95% confidence interval for this HR is given by (0.9820^{20}, 1.1236^{20}) = (0.695, 10.286)

Practice Problem (5)



The two-sided *P*-value will be based on $Z = \hat{\beta}_1$ / s.e.$(\hat{\beta}_1)$. Here $\hat{\beta}_1$= log(1.0504) = 0.0492, and

$\exp(\hat{\beta}_1 + 1.96$ s.e.$(\hat{\beta}_1))$ = 1.1236 => s.e.$(\hat{\beta}_1)$ = (log(1.1236) –log(1.0504)) / 1.96 = 0.0344,

Z = 0.0492 / 0.0344 = 1.43.

- 1.43 < 1.96 then not significant

Practice Problem (5)

- The first step with missing blood pressures would be to compare the distributions of survival times, smoking, and cholesterol values of the 25 subjects with blood pressure vs. the 10 subjects without blood pressure.
  - This could be graphically or numerically (e.g., Kaplan-Meier estimates for the survival times, and descriptive statistics, boxplots, or histograms of smoking and cholesterol), perhaps with an appropriate P-values comparing the two groups.

- Dropping 10 subjects could lead to a loss of efficiency and/or to bias, and these analyses would help us to think about the type of missingness that could be occurring here (MCAR, MAR, or NMAR), which could help us with multiple imputation or other approaches to handle the missingness appropriately (though 25 and 10 are still small sample sizes).

Next Steps in BST courses

- BST 223 (spring) Applied Survival Analysis – provides a much more extensive focus on survival modeling

- BST 226 (spring) Applied Longitudinal Analysis – provides more on correlated and longitudinal outcomes (a lot on continuous outcomes, extending linear regression, but also binary and other outcomes) and missing data

- BST 222 (fall) Basics of Statistical Inference – provides more details on where estimators and standard errors come from, the basics of maximum likelihood theory, and application to many problems in estimation, hypothesis testing, confidence intervals, and regression; requires calculus and some linear algebra

- Others in BST (and EPI and elsewhere)

- Stop by and ask me questions…

Next Steps in software

- Different courses may use (require) different software packages than you have been using in BST 210

- Possibly use January to review data input, data manipulation, and regression methods by looking at other BST 210 outputs for practice in a different software package

- Stop by and ask me questions…

Finale

- Goals of the course were to make you very comfortable with applying regression models for continuous, binary, rate, and survival outcomes, particularly from a medical/public health/ epidemiological perspective

- Model building and interpretation techniques are surprisingly similar for different types of outcomes and regression methods

- Practice on real data is important, as you are likely to analyze similar data in the future

- You can always learn more methods, extensions of methods…and develop methods yourselves ☺

- Good luck with future modeling!

Save the Date!

**** December 19: 11:30am-1:00pm FXB G12 ****

Special Program for our final class!

featuring guest speakers and their seminal
contributions to the field of statistics and modeling

And of course a celebration!

**** We expect *ALL BST 210 students* to attend. ****

(no missing data, please)

Thursday, December 19 - Special Guests

Nan Laird

Francesca Dominici

Briana Stephenson

Thank you

I USUALLY BREAK THE ICE WITH A HUMOROUS EQUATION...
3/31 DEERING © 2010 CREATORS.COM

Extremely proud of each of you – it has been such an honor to teach all of you this fall!