Final Exam Practice Questions

EPID 785R: Applied Foundations of Study Design and Regression Modeling, Fall 2026

Ashley I Naimi

What this is. Twenty-five practice questions, one per page, in the form of the pop quizzes: one topic, one small table or one fitted model, two or three short parts. They cover the topics in the final exam study guide; the table below gives each question's week and exam topic numbers. Work them closed book and by hand. Where a question prints an equation or a value of exp, log or a square root, use it; where it prints none, the definitions are part of what the exam expects you to know.

Conventions. The exposure is X (1 = exposed), the outcome is Y, covariates are C or Z. "Risk" means cumulative incidence over the stated window. A 95% interval is estimate $\pm 1.96\times$ SE throughout. The minutes are a rough guide.

#	Week	Exam topics	Min
1	1	1, 3	9
2	1	2	8
3	2	18	7
4	3	4	7
5	3	5	6
6	3	4	7
7	4	6	8
8	5	7	8
9	5	7	6
10	5	8	7
11	6	19, 11	6
12	7	11, 7	5
13	7	11	6
14	8	9	9
15	8	9, 10	6
16	9	10	7
17	10	12	5
18	10	8, 12	5
19	11	14	5
20	12	13, 19	7
21	12	13	6
22	11	12	6
23	13	15	10
24	13	16	8
25	13	17	7

About 170 minutes of work in all.

Question 1

Week 1. Exam topics 1 and 3. About 9 minutes.

An insurance-claims study emulates a target trial of starting drug D after a new diagnosis of condition E, with all-cause mortality within one year as the outcome. Every patient enters at diagnosis (day 0). The exposure was defined from the pharmacy record as *on drug D at the six-month visit (day 180)*: patients with an activeprescription for D on day 180 were analyzed as the treated arm and all other patients as the untreated arm. Follow-up for mortality ran from day 0 to day 365 in both arms.

The table shows what was observed.

Arm, as analyzed	Patients	Deaths, days 0 to 180	Deaths, days 181 to 365	Deaths by one year
On D at day 180	500	0	50	50
Not on D at day 180	500	50	63	113

(a) Which component of the protocol was assessed with information from after time zero? Name the rule it breaks, and state the direction of the bias in the as-analyzed risk difference, with one sentence on why the first 180 days are the problem.

(b) Compute the one-year risk difference as analyzed. Then move time zero to day 180: using only the patients alive on day 180, compute the risk difference for death between day 181 and day 365.

(c) In one sentence each, say what the landmark analysis in (b) changed about the population and about the question.

Question 2

Week 1. Exam topic 2. About 8 minutes.

A trial randomized 800 patients after cardiac surgery to a 12-week supervised exercise program ($X = 1$) or usual care ($X = 0$), 400 per arm. $A = 1$ means the patient completed the program (at least 75% of sessions); some usual-care patients joined an equivalent community program on their own and also count as $A = 1$. The outcome Y is hospital readmission within six months. The table gives the number of patients and the number readmitted in each cell.

Assigned X	Completed A	Patients	Readmitted	Risk
1	1	300	30	0.10
1	0	100	30	0.30
0	1	100	10	0.10
0	0	300	60	0.20

Given. ITT $= P(Y{=}1 \mid X{=}1) - P(Y{=}1 \mid X{=}0)$. Naive as-treated $= P(Y{=}1 \mid A{=}1) - P(Y{=}1 \mid A{=}0)$. Naive per-protocol $= P(Y{=}1 \mid X{=}1, A{=}1) - P(Y{=}1 \mid X{=}0, A{=}0)$.

(a) Compute the three contrasts.

(b) Which contrast does randomization identify, and why? Say in one sentence what comparing the outcome risk in groups defined by A rather than by X reintroduces, naming the arrows in the trial's DAG that do it.

Question 3

Week 2. Exam topic 18. About 7 minutes.

A cohort study of 20,000 adults aged 60 and over measured alcohol intake once, by questionnaire at enrollment, and coded the exposure as *current drinker* (any alcohol in the past 30 days) versus *non-drinker* (no alcohol in the past 30 days). Participants were followed for 15 years for incident dementia. The abstract concludes: "Current drinking was associated with a 20% lower rate of dementia (hazard ratio 0.80, 95% CI 0.70 to 0.91). These findings support moderate alcohol consumption as a protective behavior in older adults."

(a) State the intervention implied by this exposure coding as precisely as the coding allows, and list what it leaves unspecified.

(b) Describe who the referent group ("non-drinker") can contain in this population. In which direction would that composition move the hazard ratio?

(c) Say what a single baseline measurement can and cannot tell you about timing (prevalent versus newly started drinking; sustained versus one-time drinking), and name one thing the study should have collected or done at enrollment to repair the comparison.

Question 4

Week 3. Exam topic 4. About 7 minutes.

A closed cohort of 2,000 people was followed for two years with no loss to follow-up. The table gives, by exposure, the number of people at baseline, the number who became cases, and the person-time they accumulated.

Exposure	People at baseline	Cases	Person-years
Exposed	1,000	200	1,800
Unexposed	1,000	100	1,900

Three investigators then drew three control series from this same cohort:

- Series I: 10% of the people who were still non-cases at the end of follow-up: 80 exposed, 90 unexposed.
- Series II: 10% of the baseline cohort, drawn at enrollment regardless of what happened to them later: 100 exposed, 100 unexposed.
- Series III: one control per 10 person-years of follow-up, drawn from those at risk as the cases occurred: 180 exposed, 190 unexposed.

(a) From the full cohort, compute the risk ratio, the odds ratio, and the rate ratio.

(b) For each of the different control series, compute the case-control odds ratio. Match each result to one of the three cohort quantities in (a) and name the sampling scheme that produced it.

Question 5

Week 3. Exam topic 5. About 6 minutes.

A cohort of $N = 5{,}000$ people was followed for one year. To save on measuring the exposure, the investigators used a case-cohort design: a random subcohort of 10% of the cohort was drawn at baseline, and the exposure was measured on everyone in the subcohort and on every case in the full cohort. The table gives what was measured.

	Exposed	Unexposed
Subcohort members	100	400
Cases among the subcohort members	6	8
Cases in the full cohort	60	80

(a) Rebuild the weighted 2×2 table (weighted people and cases, by exposure) and run the design check: what should the weighted people sum to, and do they?

(b) Compute the two weighted risks (in the exposed and unexposed), the risk ratio, and the risk difference.

Question 6

Week 3. Exam topic 4. About 7 minutes.

In a source population of 1,000,000 people followed for one year, 4,000 developed the outcome. A case-control study enrolled every case (sampling fraction $\pi_1 = 1$) and a 1% random sample of the non-cases ($\pi_0 = 0.01$). Logistic regression of case status on a binary exposure X in the case-control data gave

$$\operatorname{logit} P(\text{case} \mid X) = -1.0 + 0.7\, X.$$

Given. $\beta_0(\text{population}) = \beta_0(\text{sample}) - \log(\pi_1/\pi_0)$. $P = \exp(\eta)/\{1 + \exp(\eta)\}$. $\log(100) = 4.6$; $\exp(0.7) = 2.0$.

(a) Compute the intercept the same model would have in the source population.

(b) Does $\exp(0.7)$ from the fitted model need a correction? Why or why not?

Question 7

Week 4. Exam topic 6. About 8 minutes.

A hospital fit a logistic regression of 30-day readmission on enrollment in a new discharge-planning program (X), age, sex and diabetes, in 4,000 discharged patients. The program was offered preferentially to patients with diabetes, and the prevalence of readmission in the data was roughly 30%. The coefficients for the model's fit were:

Term	Estimate	SE	p
Intercept	-0.80	0.07	< 0.001
Program (X)	-0.40	0.12	0.001
Age (per 10 years, centered at 65)	0.15	0.05	0.003
Male	0.10	0.11	0.36
Diabetes	0.05	0.10	0.62

The program's director then used this fit to estimate the program's effect on readmission in their hospital.

(a) A colleague proposes dropping diabetes "because $p = 0.62$" and refitting. Describe both the (i) causal and (ii) statistical consequences that might arise as a result of dropping diabetes and refitting the model to estimate the program's effect.

(b) The draft report calls every coefficient in the table an "independent risk factor for readmission," and reports the coefficients and standard errors as "effects." Name the fallacy entailed by this practice, and describe the specific problems with doing so.

Question 8

Week 5. Exam topic 7. About 8 minutes.

A Poisson regression of the number of hospitalizations Y over each person's follow-up time T (in years) on an exposure X (0/1) and age gave

$$\log E(Y \mid X, \text{age}, T) = -4.0 + 0.70\,X + 0.03\,(\text{age} - 50) + \log T,$$

where the coefficient of $\log T$ was fixed at 1 rather than estimated.

Given. $\exp(0.70) = 2.0$; $\exp(-3.0) = 0.050$. Expected count = rate $\times T$.

(a) Label the five parts of this model: link, distribution, target function, nuisance function, and offset.

(b) What does $\exp(0.70)$ report? Provide a full interpretation, including the units of the contrast.

(c) For an exposed 60-year-old followed for two years, compute the expected hospitalization rate (per person-year) and the expected number of hospitalizations. At what age does the age term contribute nothing?

(d) In one sentence, what would $\exp(0.70)$ report if $\log T$ were left out of the model entirely?

Question 9

Week 5. Exam topic 7. About 6 minutes. (Pairs with Question 12.)

A logistic regression of an outcome Y on an exposure X (0/1), age, sex ($F = 1$ for women) and an exposure-by-sex product term gave

$$\text{logit}\, P(Y{=}1) = -1.5 + 0.8\, X - 0.04\,(\text{age} - 60) + 0.5\, F - 0.3\, XF.$$

(a) Compute the predicted probability of the outcome for an exposed woman aged 70.

(b) Calculate the odds ratio for both men and women using the coefficients from the model.

(c) Say in one sentence what the product-term coefficient of -0.3 measures. In other words, provide an interpretation of exp(-0.3) from this model.

Question 10

Week 5. Exam topic 8. About 7 minutes.

The table gives a binary outcome by binary exposure X and binary covariate C status, with 100 people in every cell.

C	X	People	Cases	Risk
0	1	100	20	0.20
0	0	100	10	0.10
1	1	100	60	0.60
1	0	100	30	0.30

Given. Three additive models with three parameters each: identity link, $P = \alpha_0 + \alpha_1 X + \alpha_2 C$; log link, $\log P = \gamma_0 + \gamma_1 X + \gamma_2 C$; logit link, $\operatorname{logit} P = \delta_0 + \delta_1 X + \delta_2 C$. $\log(0.1) = -2.30$; $\log 2 = 0.69$; $\log 3 = 1.10$; $\exp(-0.51) = 0.60$.

(a) Within each stratum of C, compute the RD, the RR and the OR. Which measure is the same in both strata?

(b) Which of the three additive models reproduces all four cell risks exactly? Write its three coefficients using the printed logs, and show that its fitted risk for $X = 1, C = 1$ equals the table.

Question 11

Week 6. Exam topics 19 and 11. About 6 minutes.

In a cohort of 200 people with a binary outcome (100 exposed with risk 0.20; 100 unexposed with risk 0.10), a log-binomial model of the risk ratio failed to converge, so the analyst fit a Poisson model with a log link to the same 0/1 outcome and reported $\exp(\beta_1) = 2.0$ with a model-based standard error.

Given. Variance functions: Gaussian, constant; binomial, $\mu(1-\mu)$; Poisson, μ; gamma, μ^2. Approximate variance of the log of a group's risk, for a group of n people with risk μ: binomial, $(1-\mu)/(n\mu)$; Poisson, $1/(n\mu)$. $\sqrt{1.25} = 1.12$.

(a) For the exposed group ($n = 100$, $\mu = 0.20$), compute the variance of the log risk under the binomial family and under the Poisson family, and the ratio of the two standard errors (Poisson to binomial).

(b) In which direction is the Poisson model-based standard error wrong, for any risk strictly between 0 and 1? Name what repairs it without changing the coefficient.

Question 12

Week 7. Exam topics 11 and 7. About 5 minutes. (Pairs with Question 9.)

Return to the logistic model of Question 9, $\operatorname{logit} P(Y{=}1) = -1.5 + 0.8\,X - 0.04\,(\text{age} - 60) + 0.5\,F - 0.3\,XF$. The part of its estimated variance-covariance matrix for the coefficients of X, F and XF is:

	X	F	XF
X	0.040	-0.010	-0.035
F	-0.010	0.050	-0.030
XF	-0.035	-0.030	0.090

Given. $\operatorname{Var}(A + B) = \operatorname{Var}(A) + \operatorname{Var}(B) + 2\operatorname{Cov}(A, B)$. $\operatorname{Var}(A + wB) = \operatorname{Var}(A) + w^2 \operatorname{Var}(B) + 2w \operatorname{Cov}(A, B)$. $\sqrt{0.060} = 0.245$; $\sqrt{0.130} = 0.361$; $\exp(0.02) = 1.02$; $\exp(0.98) = 2.66$.

(a) Compute the variance and the standard error of the exposure log odds ratio among women, $\beta_X + \beta_{XF}$.

(b) Compute its 95% Wald confidence interval, and the interval for the odds ratio.

Question 13

Week 7. Exam topic 11. About 6 minutes.

A study of 60 primary-care clinics (20 patients each, 1,200 patients) compared systolic blood pressure after one year between the 30 clinics that adopted a new hypertension protocol and the 30 that did not; adoption was a clinic-level decision. A linear model of blood pressure on the protocol indicator and patient covariates gave a coefficient of -3.0 mmHg. Four standard errors were computed for it.

Standard error	Value
Model-based (from the fitted family's variance function)	1.00
HC3 robust (sandwich, patients treated as independent)	1.10
Cluster-robust (sandwich, clinics as clusters)	1.50
SD of 1,000 cluster-bootstrap replicates	1.55

Given. 95% interval = estimate $\pm 1.96\times$ SE.

(a) In one sentence each, say what the four standard errors are built from. Which one should be reported here, and what counts as the sample size for it?

(b) Compute the 95% confidence interval from the standard error you chose.

Question 14

Week 8. Exam topic 9. About 9 minutes.

Z is a cause of the outcome but is unrelated to the exposure: half of each exposure group has $Z = 1$. There are 100 people in every cell.

Z	X	People	Cases	Risk
0	1	100	50	0.50
0	0	100	20	0.20
1	1	100	80	0.80
1	0	100	50	0.50

(a) Compute the RD, the RR and the OR within each stratum of Z, and in the whole table ignoring Z.

(b) Show the arithmetic that recovers the overall RR from the four stratum risks

(c) Is there a comparable arithmetic (i.e., weighted sum) that can recover the overall OR from the stratum specific ORs? Why or why not?

Question 15

Week 8. Exam topics 9 and 10. About 6 minutes.

A logistic regression of a binary outcome on the exposure X and a binary covariate Z gave

$$\operatorname{logit} P(Y{=}1 \mid X, Z) = -1.39 + 1.39\,X + 1.39\,Z.$$

In the study population 30% of people have $Z = 1$, and Z is the only confounder.

Given. $P = \exp(\eta)/\{1+\exp(\eta)\}$. $\exp(-1.39) = 0.25$; $\exp(0) = 1$; $\exp(1.39) = 4.0$. Standardized risk under $X = x$: $\sum_z P(Y{=}1 \mid X{=}x, Z{=}z)\,P(Z{=}z)$.

(a) Compute the four fitted risks, then the standardized risk if everyone were exposed and if no one were exposed.

(b) Compute the standardized (marginal) RD, RR and OR, and compare the OR with exp(1.39).

Question 16

Week 9. Exam topic 10. About 7 minutes.

A cohort of 200 people, with a binary confounder C and a binary outcome.

C	X	People	Cases	Risk
0	1	20	4	0.20
0	0	80	8	0.10
1	1	60	30	0.50
1	0	40	16	0.40

Given. For a person with $X = x$ and $C = c$: unstabilized weight $w = 1/P(X{=}x \mid C{=}c)$; stabilized weight $sw = P(X{=}x)/P(X{=}x \mid C{=}c)$. Weighted risk in arm x = (sum of weight $\times$ cases) $\div$ (sum of weight $\times$ people) over the cells with $X = x$.

(a) Compute $P(X{=}1 \mid C{=}0)$, $P(X{=}1 \mid C{=}1)$, $P(X{=}1)$, and the unstabilized and stabilized weights for each of the four cells.

(b) Compute the total weighted population under each set of weights and, under the stabilized weights, the proportion exposed within each level of C. State what each of these three numbers checks.

Question 17

Week 10. Exam topic 12. About 5 minutes.

A linear spline (piecewise linear function) with knots at 2 and 5 was fit for the mean of a continuous outcome Y as a function of a continuous exposure x (0 to 8 units):

$$E(Y \mid x) = 10 + 3\,x - 4\,(x-2)_+ + 2\,(x-5)_+,$$

where $(x-\kappa)_+$ equals $x-\kappa$ when $x > \kappa$ and 0 otherwise.

(a) Compute the predicted mean at $x = 1$ and at $x = 6$, and the contrast between them.

(b) Say why a spline fit is generally reported as predictions and contrasts of those predictions at chosen values of x, rather than as its coefficients.

Question 18

Week 10. Exam topics 8 and 12. About 5 minutes.

Three logistic models of the same outcome on a binary exposure X and one continuous confounder C (60 distinct values) were fit to the same people. They differ only in how C enters: linearly; as a spline with 4 degrees of freedom; as one indicator (i.e., dummy variable) per distinct value of C.

Model for C	Parameters p	Deviance	Exposure coefficient (SE)
Linear	3	950.0	0.60 (0.20)
Spline, 4 df	6	940.0	0.45 (0.21)
One indicator per value	61	900.0	0.44 (0.24)

(a) Rank each of the three fits along the "parametric-to-nonparametric" spectrum, and say which end carries less bias and which less variance.

(b) The question is causal: the effect of X on the outcome. Choose a form for C as the model from which you would report the effect of X. Justify your choice in one or two sentences.

Question 19

Week 11. Exam topic 14. About 5 minutes.

Four candidate models for a nuisance function $g(C)$ in $E(Y \mid X, C) = \beta X + g(C)$ were fit to 2,000 people. The table gives each candidate's in-sample mean squared error (MSE) and its 5-fold cross-validated MSE.

Candidate for $g(C)$	In-sample MSE	5-fold CV MSE
Linear in C	25.0	25.4
Spline, 3 df	24.0	24.3
Spline, 8 df	23.0	24.9
One indicator per value of C	20.0	28.0

A super learner fit on the same folds put weight 0.70 on the 3-df spline and 0.30 on the linear model (0 on the other two). For one new person, the 3-df spline predicts 4.0 and the linear model predicts 5.0.

(a) Which candidate does in-sample loss pick, and which does cross-validation pick? Which is more valid?

(b) Compute the non-discrete super learner's prediction for the new person, and show how you calculate it. Name the algorithm that the "discrete super learner" would pick among the four candidates.

Question 20

Week 12. Exam topics 13 and 19. About 7 minutes.

	Cases	Non-cases	Total
Exposed	7	0	7
Unexposed	20	20	40

Logistic regression of case status on exposure returns an exposure coefficient of 19.6 with a standard error of 2,900.

(a) Say why the odds ratio for this table is not a number. How, exactly, does the maximum-likelihood fitting loop from the logistic model produce an estimate of 19.6?

(b) Use Firth's method and add half a person to each cell, and compute the odds ratio in this sparse data table. Then, suppose the exposed group had instead been 80 total people with 70 cases (the unexposed row unchanged). Compute the odds ratio for this new table, and also use Firth's method to compute the "half-person" odds ratio in this new table. Compare the two odds ratios from the two tables (the fit's coefficient of 19.6, an OR of exp(19.6), and Firth's OR in the sparse table; the OR and Firth's OR in the unsparse table). Say in one sentence what happens to the half-person's pull as the size of the cells in the table increase.

(c) Name the two remedies for the empty cell discussed in the course. For each remedy, discuss whether it changes the question being asked, or the estimator being used, or both.

Question 21

Week 12. Exam topic 13. About 6 minutes.

An analyst wants the effect of an exposure X (0/1) on a continuous outcome Y, adjusting for three confounders C_1, C_2, C_3, all known to affect both X and Y. Two fits to 1,000 people:

Term	OLS estimate (SE)	Lasso at the CV-chosen λ
X	3.0 (0.8)	2.1
C_1	2.0 (0.5)	1.5
C_2	0.4 (0.4)	0.0
C_3	-1.0 (0.5)	-0.7

C_2 is strongly associated with X. The analyst reports the lasso's coefficient for X with a 95% interval of 1.4 to 2.8, obtained by rerunning the lasso on 2000 bootstrap samples and taking the 2.5th and 97.5th percentiles of the 2000 coefficients.

(a) Name the two things the lasso did to the coefficient of X, and explain why the second one leaves confounding in the estimate.

(b) Say what the bootstrap interval ignores, and what that can do to its coverage.

Question 22

Week 11. Exam topic 12. About 6 minutes.

An analyst wants the effect of a binary exposure A on a continuous outcome Y, adjusting for a set of confounders W, with the partially linear model $E(Y \mid A, W) = \psi A + g(W)$, where g is left unspecified. Two nuisance models were fit with a super learner and 5-fold cross-fitting: the outcome model $E(Y \mid W)$ and the exposure model $E(A \mid W)$. The table gives, for four people from one held-out fold, the exposure, the outcome, and the two out-of-fold predictions.

Person	A	$\hat{E}(A \mid W)$	Y	$\hat{E}(Y \mid W)$
1	1	0.5	10.0	8.0
2	0	0.5	7.0	8.0
3	1	0.2	12.4	10.0
4	0	0.2	9.4	10.0

Given. Exposure residual $\tilde{A} = A - \hat{E}(A \mid W)$; outcome residual $\tilde{Y} = Y - \hat{E}(Y \mid W)$; residual-on-residual estimate (least squares through the origin) $\hat{\psi} = \sum_i \tilde{A}_i \tilde{Y}_i \big/ \sum_i \tilde{A}_i^2$.

(a) Compute the two residuals for each person and the estimate $\hat{\psi}$, treating the four people as the whole sample.

(b) The predictions in the table come from models fit on the other four folds. Why? What goes wrong if each person's own data are used to fit them?

(c) In one sentence, what does $\hat{\psi}$ estimate, and what must hold, beyond the identification conditions, for it to be the average treatment effect?

Question 23

Week 13. Exam topic 15. About 10 minutes.

Eight patients were followed from kidney transplant. Event A is graft failure; event B is death with a functioning graft; "censored" means the patient left the study alive with a functioning graft.

Patient	Months	Status
1	1	A, graft failure
2	2	censored
3	3	A, graft failure
4	4	B, death
5	5	A, graft failure
6	6	censored
7	7	B, death
8	8	A, graft failure

Given. Kaplan-Meier all-cause survival: $S(t) = \prod_{t_k \le t}(1 - d_k/n_k)$, the product over the distinct event times t_k up to t, where d_k is the number of events of either type at t_k and n_k the number still at risk just before t_k. Aalen-Johansen risk of cause j: $F_j(t) = \sum_{t_k \le t} S(t_k^-)\, d_{jk}/n_k$, where $S(t_k^-)$ is the Kaplan-Meier all-cause survival just before t_k and d_{jk} the number of cause-j events at t_k.

(a) Before any estimator: using counts alone, give the lowest and the highest value the all-cause risk at 5 months could take, depending on what happened to the patient censored before month 5. Then compute $1 - S(5)$ by Kaplan-Meier and check that it lies between them.

(b) Compute the Aalen-Johansen risk of graft failure (A) at 5 months and of death (B) at 5 months, and check that they sum to $1 - S(5)$.

(c) An analyst codes the deaths (B) as censored and computes a Kaplan-Meier "risk of graft failure" at 5 months. Compute it, say by how much it exceeds the Aalen-Johansen value, and say in one sentence what that excess represents.

Question 24

Week 13. Exam topic 16. About 8 minutes.

A manuscript says: "Because the hazard ratio is a biased estimate of the causal effect in heterogeneous populations, we report the risk difference at five years instead." For the sake of the question, suppose we knew both potential survival times for all 100 patients in the study:

Would survive to 5 years under …	Patients
both treatment and control	50
control only	20
treatment only	5
neither	25

(a) Say what is wrong with the sentence, in estimand-versus-estimator language, in two sentences.

(b) The course gave two properties that a causal estimand should have. State them. Then count from the table: how many patients are in the denominator of the hazard at 5 years under treatment, how many under control, and how many in the denominator of each arm's 5-year risk? Compute the causal risk difference at 5 years.

(c) Say why the hazard ratio's selection problem grows over follow-up in a heterogeneous population even when every hazard is estimated without error, and name the row of the table that causes it.

Question 25

Week 13. Exam topic 17. About 7 minutes.

Two models were fit to the same data on time to a first cardiovascular event, with a binary exposure X. The Cox model gave $\beta = 0.69$ for X, and the baseline survival for the unexposed at 5 years was estimated afterwards as $S_0(5) = 0.80$. The accelerated failure time (AFT) model gave $\gamma = -0.69$ for X, meaning that the exposed reach any given survival probability in half the time of the unexposed; its fitted survival curve for the unexposed gives $S_0(5) = 0.80$ and $S_0(10) = 0.60$.

Given. Cox: $S(t \mid X{=}1) = S_0(t)^{\exp(\beta)}$. AFT: $S(t \mid X{=}1) = S_0\{t \times \exp(-\gamma)\}$. $\exp(0.69) = 2.0$; $\exp(-0.69) = 0.50$. Risk $= 1 - S$.

(a) From the Cox fit, compute the 5-year risk for an exposed person and for an unexposed person.

(b) From the AFT fit, compute the 5-year risk for an exposed person.

(c) Say what the parametric baseline buys and what it costs, and why the Cox fit alone reports hazard ratios and not risks: what has to be estimated afterwards, and why it is a nuisance function.