Exam 3 is Wed. April 26. **You are allowed 13 sheets of notes and a calculator.** The exam covers HW7-10 and Q7-9. Numbers refer to types of problems on exam. See the last page for the final: Tuesday, May 9, 2:45-4:45.

If there are important predictors such as treatment that must be in the submodel, either force the variable selection procedures to contain the important predictors or do variable selection on the less important predictors and then add the important predictors to the submodel.

A **scatterplot** is a plot of x_i vs. x_j. A **scatterplot matrix** is an array of scatterplots. It is used to examine the marginal relationships of the predictors. Variables with outliers, missing values or strong nonlinearities may be so bad that they should not be included in the full model.

36) Suppose that all values of the variable x are positive. The **log rule** says add $\log(x)$ to the full model if $\max(x_i)/\min(x_i) > 10$.

37) Suppose the PH model contains $x_1, ..., x_p$. Leave out x_j, find the martingale residuals $r_{m(j)}$, plot x_j vs $r_{m(j)}$ and add the lowess or loess curve. If the curve is linear then x_j has the correct funtional form. If the curve looks like $t(x_j)$ (eg $(x_j)^2$), then replace x_j by $t(x_j)$, find the martingale residuals, plot $t(x_j)$ vs the residuals and check that the loess curve is linear. See HW7 1ab.

38) Let the scaled Schoenfeld residual for the jth variable x_j be $r^*_{pj} + \hat{\beta}_j$. Plot the death times t_i vs the scaled residuals and add the loess curve. If the loess curve is approximately horizontal for each of the p plots, then the PH assumption is reasonable. Alternatively, fit a line to each plot and test that each of the p slopes is equal to 0. The R function `cox.zph` makes both the plots and tests. See HW7 1cd. For the output below, the PH assumption is reasonable since the Global pval $= 0.349 > \delta = 0.05$. If the Global pvalue $< \delta = 0.05$, then the PH assumption is unreasonable.

```
cox.zph(lung.fit2)
           rho      chisq   p
pph.ecog    0.05189 0.3905 0.532
ph.karno   0.14216  2.2081 0.137
pat.karno 0.04773  0.3812 0.537
wt.loss    0.00857  0.0131 0.909
GLOBAL          NA  4.4476 0.349
```

39) Suppose the observed survival times $T_1, ..., T_n$ are a censored data set from an Exponential (λ) distribution. Let $T_i = Y_i^*$. Let $\delta_i = 0$ if the case is censored and let $\delta_i = 1$, otherwise. Let $r = \sum_{i=1}^n \delta_i =$ the number of uncensored cases. Then the MLE $\hat{\lambda} = r/\sum_{i=1}^n T_i$. So $\hat{\lambda} = r/\sum_{i=1}^n Y_i^*$. A 95% CI for λ is $\hat{\lambda} \pm 1.96\hat{\lambda}/\sqrt{r}$. See HW8 1.

40) The **Weibull proportional hazards regression (WPH) model** is

$$h_i(t) = h_{Y_i|\boldsymbol{x}_i}(t) = h_{Y_i|\boldsymbol{\beta}'_W\boldsymbol{x}_i}(t) = \exp(\boldsymbol{\beta}'_W\boldsymbol{x}_i)h_0(t)$$

where $h_0(t) = h_0(t|\boldsymbol{\theta}) = \lambda\gamma t^{\gamma-1}$ is the **baseline function**. So $Y|SP \sim W(\lambda\exp(SP), \gamma)$.

For now, assume that the WPH model is appropriate, although this assumption should be checked before performing inference.

Shown below is output in symbols from *SAS* and *R* . The estimated coefficient is $\hat{\beta}_j$. The Wald chi square $= X^2_{o,j}$ while p and "pr > chisqu" are both p-values.

For SAS only.
log likelihood log L(none)

variable	Estimate	Std. Error	Est/SE	or $(Est/SE)^2$	p-value
intercept					
scale					
Weibull shape					

For SAS or R

variable	Estimate	Std. Error	Est/SE	or $(Est/SE)^2$	p-value
intercept					
x_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	$z_{o,1} = \hat{\beta}_1/se(\hat{\beta}_1)$	$X^2_{o,1} = z^2_{o,1}$	for Ho: $\beta_1 = 0$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	$\hat{\beta}_p$	$se(\hat{\beta}_p)$	$z_{o,p} = \hat{\beta}_p/se(\hat{\beta}_p)$	$X^2_{o,p} = z^2_{o,p}$	for Ho: $\beta_p = 0$
scale or	log scale				
Weibull shape	or scale				

For the full model, SAS will have Log Likelihood = log L(full).
For the full model, R will have log L(full), log L (none) and
chisq = [-2 log L(none)] - [-2 log L(full)] on p degrees of freedom with pvalue
Replace full by reduced for the reduced model.

The SAS and R log likelihood, log L, differ by a constant.

```
SAS Log Likelihood = -29.7672 null model
variable      df Estimate  SE     chi square  pr > chisqu
intercept     1   7.1110   0.2927  590.12       < 0.0001
Weibull Scale 1   1225.4   358.7
Weibull Shape 1   1.1081   0.2810

SAS Log Likelihood = -29.1775  reduced model
variable      df Estimate  SE     chi square  pr > chisqu
intercept     1   7.3838   0.4370 285.45         < 0.0001
treat         1  -0.5593   0.5292   1.12           0.2906
Scale         1   0.8857   0.2227
Weibull Shape 1   1.1291   0.2840

SAS Log Likelihood = -20.5631    full model
variable      df Estimate  SE     chi square  pr > chisqu
intercept     1  11.5483   1.1970 93.07         < 0.0001
age           1  -0.0790   0.0198 15.97         < 0.0001
```

```
treat         1  -0.5615   0.3399 2.73           0.0986
Scale         1   0.5489   0.1291
Weibull Shape 1   1.8218   0.4286

R reduced  model Value Std. Error      z        p
(Intercept)      7.384      0.437 16.895 4.87e-64
treat           -0.559      0.529 -1.057 2.91e-01
Log(scale)      -0.121      0.251 -0.483 6.29e-01
Scale= 0.886
Loglik(model)= -97.4   Loglik(intercept only)= -98
        Chisq= 1.18 on 1 degrees of freedom, p= 0.28

R full model     Value Std. Error     z        p
(Intercept)     11.548     1.1970  9.65 5.04e-22
treat           -0.561     0.3399 -1.65 9.86e-02
age             -0.079     0.0198 -4.00 6.43e-05
Log(scale)      -0.600     0.2353 -2.55 1.08e-02
Scale= 0.549
Loglik(model)= -88.7   Loglik(intercept only)= -98
        Chisq= 18.41 on 2 degrees of freedom, p= 1e-04
```

41) Instead of fitting the WHP model of 40), R and SAS fit an accelerated failure time model $\log(Y_i) = \alpha + \boldsymbol{\beta}'\boldsymbol{x}_i + \sigma\epsilon_i$ where $\text{Var}(\epsilon_i) = 1$ and the ϵ_i are iid from a smallest extreme value distribution. Also $\boldsymbol{\beta} \neq \boldsymbol{\beta}_W$ from 40). Then intercept corresponds to α and scale to σ. Hence intercept, shape, Weibull shape, scale and log(scale) **are not predictors** $x_1, ..., x_p$.

$\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ are MLEs found from the censored data $(T_i, \delta_i, \boldsymbol{x}_i)$ not from $(Y_i, \boldsymbol{x}_i)$.

42) Let $\log(T_i) = \hat{\alpha} + \hat{\boldsymbol{\beta}}'\boldsymbol{x}_i + r_i$. A *log censored response (LCR) plot* is a plot of $\hat{\alpha} + \hat{\boldsymbol{\beta}}'\boldsymbol{x}_i$ vs $\log(T_i)$ with plotting symbol 0 for censored cases and + for uncensored cases. The vertical deviations from the identity line $= r_i$. The least squares line based on the +'s is also added to the line and should have slope not too far from 1, especially if $\gamma \geq 1$. The plotted points should be linear with roughly constant variance. The censoring and long left tails of the smallest extreme value distribution make judging linearity and detecting outliers from the left tail difficult. Try to ignore the bottom of the plot where there are few cases when assessing linearity.

43) Given $\hat{\boldsymbol{\beta}}$ from output and given $\boldsymbol{x}$, be able to find ESP $= \hat{\boldsymbol{\beta}}'\boldsymbol{x} = \sum_{i=1}^{p} \hat{\beta}_i x_i = \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$.

44) A large sample 95% CI for β_j is $\hat{\beta}_j \pm 1.96\ se(\hat{\beta}_j)$.

45) **4 step Wald test of hypotheses**:

i) State the hypotheses Ho: $\beta_j = 0$ Ha: $\beta_j \neq 0$.

ii) Find the test statistic $z_{o,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or $X^2_{o,j} = z^2_{o,j}$ or obtain it from output.

iii) The p–value $= 2P(Z < -|z_{oj}|) = P(\chi^2_1 > X^2_{o,j})$. Find the p–value from output or use the standard normal table.

iv) If pval $< \delta$, reject Ho and conclude that x_j is needed in the Weibull survival model given that the other $p-1$ predictors are in the model. If pval $\geq \delta$, fail to reject Ho, and conclude that the values of x_j do not affect the WPH survival model given that the other $p-1$ predictors are in the model. (Or state that there is not enough evidence to conclude that the values of x_j affect the WPH survival model.)

46) The 4 step likelihood ratio test **LRT** is

i) $H_o : \boldsymbol{\beta} = \mathbf{0}$ $H_A : \boldsymbol{\beta} \neq \mathbf{0}$

ii) test statistic $X^2(N|F) = [-2 \log L(none)] - [-2 \log L(full)]$ is often obtained from output

iii) The p–value $= P(\chi^2_p > X^2(N|F))$ where χ^2_p has a chi–square distribution with p degrees of freedom. The p–value is often obtained from output.

iv) Reject H_0 if the p–value $< \delta$ and conclude that there is a WPH survival relationship between Y and the predictors $\boldsymbol{x}$. If p–value $\geq \delta$, then fail to reject H_0, and conclude that the values of the predictors $\boldsymbol{x}$ do not (significantly) affect the WPH survival model. (Or state that there is not enough evidence to conclude that the values of $\boldsymbol{x}$ affect the WPH survival model.)

47) The 4 step **change in LR test** is

i) H_o: the reduced model is good H_A: use the full model

ii) test statistic $X^2(R|F) = X^2(N|F) - X^2(N|R) = [-2 \log L(red)] - [-2 \log L(full)]$

iii) The p–value $= P(\chi^2_{p-r} > X^2(R|F))$ where χ^2_{p-r} has a chi–square distribution with $p-r$ degrees of freedom.

iv) Reject H_o if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_o and conclude that the reduced model is good (the values of $\boldsymbol{x}_O$ do not affect the survival model, or there is not enough evidence to conclude that the values of $\boldsymbol{x}_O$ affect the survival model).

48) R and SAS programs do not have a variable selection option, but the WPH model is a PH model, so use SAS Cox PH variable selection to suggest good submodels. Then fit each candidate with WPH software and check the WPH assumptions.

49) The **accelerated failure time (AFT) model** has $\log(Y_i) = \alpha + \boldsymbol{\beta}' \boldsymbol{x}_i + \sigma \epsilon_i$ where $\text{Var}(\epsilon_i) = 1$ and the ϵ_i are iid from a location scale family.

If the Y_i are Weibull, the ϵ_i are from a smallest extreme value distribution. The Weibull regression model is both a proportional hazards model and an accelerated failure time model.

If the Y_i are lognormal, the ϵ_i are normal.

If the Y_i are loglogistic, the ϵ_i are logistic.

50) Still use the *log censored response (LCR) plot* of 42). The LCR plot is easier to use when the ϵ_i are normal or logistic since these are symmetric distributions.

51) For the AFT model, $h_i(t) = e^{-SP} h_o(t/e^{SP})$ and $S_i(t) = S_0(t/\exp(SP))$.

52) Inference for the AFT model is performed exactly in the same way as for the WPH = Weibull AFT. See points 43) – 47). But the conclusion change slightly if the AFT is not the Weibull AFT. In point 45, change (if necessary) "Weibull survival model" to the appropriate model, eg "lognormal survival model". In point 46, change (if necessary) "WPH" to the appropriate model, eg "lognormal AFT".

In principle, the slice survival plot can be made for parametric AFT models, but the programming may be difficult.

The loglogistic and lognormal AFT models are not PH models. The loglogistic AFT is a proportional odds model.

53) Let $\boldsymbol{\beta}_C$ correspond to the Cox regression and $\boldsymbol{\beta}_A$ correspond to the AFT. An EE plot is a plot of the parametric ESP vs a semiparamtric ESP with the identity line added as a visual aid. The plotted points should follow the identity line with a correlation tending to 1.0 as $n \to \infty$.

54) For the Exponential regression model, $\sigma = 1$, and $\boldsymbol{\beta}_C = -\boldsymbol{\beta}_A$. The Exponential EE plot is a plot of $-ESPE = -\hat{\boldsymbol{\beta}}_A' \boldsymbol{x}$ vs $ESPC = \hat{\boldsymbol{\beta}}_C' \boldsymbol{x}$.

55) For the Weibull regression model, $\sigma = 1$, and $\boldsymbol{\beta}_C = -\boldsymbol{\beta}_A/\sigma$. The Weibull EE plot is a plot of

$$-ESPW/\hat{\sigma} = -\frac{1}{\hat{\sigma}} \hat{\boldsymbol{\beta}}_A' \boldsymbol{x} \text{ vs ESPC} = \hat{\boldsymbol{\beta}}_C' \boldsymbol{x}.$$

56) The **stratified proportional hazards regression (SPH) model** is

$$h_{\boldsymbol{x},j}(t) = h_{Y_i|\boldsymbol{x},j}(t) = h_{Y_i|\boldsymbol{\beta}'\boldsymbol{x}_i}(t) = \exp(\boldsymbol{\beta}'\boldsymbol{x}_i) h_{0,j}(t)$$

where $h_{0,j}(t)$ is the **unknown baseline function** for the jth stratum, $j = 1, ..., J$ where $J \geq 2$.

A SPH model is not a PH model, but a PH model is fit to each of the J strata. The same $\boldsymbol{\beta}$ is used for each group = stratum, but the baseline hazard functions differ. Stratification can be useful if there are clusters of cases such that the observations within the clusters are not independent. A common example is the variable *study sites* and the stratification should be on site. Sometimes stratification is done on a categroical variable such as gender.

57) Inference is done exactly as for the PH model. See points 20), 21), 22), 23), and 24). Except the conclusion is changed slightly: in 22) and 23) replace "PH" by "SPH".

58) Time dependent variables $x_i(t)$ depend on time. In the PH model, the variables are fixed: they do not depend on time. Let $\boldsymbol{x}(t) = (x_1(t), ..., x_p(t))'$ where $x_j(t) \equiv x_j(0) = x_j$ for fixed variables.

The most common time dependent variables are i) a change in treatment and ii) repeated measurements of the variable for the subject (eg monthly cholesterol level or quarterly PSA level).

59) The **generalized Cox regression (GCR)** model is

$$h_{Y|\boldsymbol{\beta}'\boldsymbol{x}(t)}(t) \equiv h_{\boldsymbol{x}(t)}(t) = \exp(\boldsymbol{\beta}'\boldsymbol{x}(t)) h_0(t).$$

The baseline function $h_0(t)$ is the hazard function for subjects who have $\boldsymbol{x}(t) \equiv \boldsymbol{x}(0) = \boldsymbol{0}$, that is, their variables are 0 at the time of origin and remain 0 through time.

60) **Inference is done exactly as for the PH model.** See points 21), 22), 23), and 24). **Except the conclusion is changed slightly**: in 22) and 23) replace "PH" by "GCR".

61) Suppose a PH model has been fit with predictors $x_1, ..., x_p$. To check the PH assumption, let the PH model be the reduced GCR model and let the full GCR model have $x_1, ..., x_p, x_1 \log(t), ..., x_p \log(t)$ where $x_i \log(t)$ is the interaction between x_i and $\log(t)$. If the "reduced model is good" (fail to reject Ho), **then the PH assumption is reasonable**. Also look at the p Wald tests for $x_i \log(t)$, but remember that if $\beta_k = 0$ for all p interactions, about $\delta p = 0.05p$ or one in twenty will be incorrectly rejected if $\delta = 0.05$. See HW11.1.

Let x_1 = treatment." There could be "no treatment effect" (fail to reject Ho: $\beta_1 = 0$) if there is a time dependent variable $x_2(t)$ that accounts for the treatment effect. Then $x_2(t)$ has masked the treatment effect. For example, if x_1 is 0 for a placebo and 1 for a leukemia medicine and $x_2(t)$ is the white blood cell count, then for $x_1 = 1$, $x_2(t)$ could quickly become high and survival is high. But for $x_1 = 0$, $x_2(t)$ and survival stay low.

62) A shortcut in R for the change in PLRT test from 24) in the Exam 2 review is to use the `anova` function.

```
full <- coxph(Surv(alung[,1],alung[,2])~perf+age+ttoent+size+type+ttype+trt,
data=alung)  #full model
red <- coxph(formula = Surv(alung[, 1], alung[, 2]) ~ perf  +
    size + ttype + trt, data = alung) #reduced model
anova(full,red)
   loglik  Chisq Df P(>|Chi|)
1 -87.608
2 -87.817 0.4189  3    0.9363
   loglik  Chisq     Df P(>|Chi|)
1
2          X^2(R|F)     pval for change in PLRT test
```

63) Consider predicting a future test value Y_f given a $p \times 1$ vector of predictors $\boldsymbol{x}_f$ and training data $(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_n)$. A large sample $100(1-\delta)\%$ *prediction interval* (PI) for Y_f has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n)$ is eventually bounded below by $1-\delta$ as the sample size $n \to \infty$. A large sample $100(1-\delta)\%$ PI is *asymptotically optimal* if it has the shortest asymptotic length: the length of $[\hat{L}_n, \hat{U}_n]$ converges to $U_s - L_s$ as $n \to \infty$ where $[L_s, U_s]$ is the population shorth: the shortest interval covering at least $100(1-\delta)\%$ of the mass.

64) Let $Z_1, ..., Z_n$ be random variables, let $Z_{(1)}, ..., Z_{(n)}$ be the order statistics, and let c be a positive integer. Compute $Z_{(c)} - Z_{(1)}, Z_{(c+1)} - Z_{(2)}, ..., Z_{(n)} - Z_{(n-c+1)}$. Let shorth(c) = $[Z_{(s)}, Z_{(s+c-1)}]$ correspond to the interval with the shortest length.

End Exam 3 Material

Below is for Quiz 10 and the Final

65) The large sample $100(1-\delta)\%$ *shorth*(c) *CI* uses the interval $[T^*_{(1)}, T^*_{(c)}], [T^*_{(2)}, T^*_{(c+1)}], ..., [T^*_{(B-c+1)}, T^*_{(B)}]$ of shortest length. Here $c = \min(B, \lceil B[1-\delta+1.12\sqrt{\delta/B}\ ]\ \rceil)$. The shorth CI is computed by applying the shorth PI to the bootstrap sample.

```
66)         Estimate Std.Err 95% shorth CI
X1         -42.4846 51.2863 [-192.281, 52.492]
X2          0               [   0.000,  0.268]
X3          1.1707  0.0598 [   0.992,  1.289]
X4          0               [   0.000,  0.840]
X5          0               [   0.000,  1.916]
X6          0.1467  0.0368 [  0.0747,  0.215]
```

Given output such as that above, be able to find a CI for β_i and $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$. Note that the CI for $\boldsymbol{\beta}_3$ is [0.992, 1.289] and $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0} = (-42.4846, 0, 1.1707, 0, 0, 0.1467)^T$.

67) Given B bootstrap samples, be able to compute the statistic T^*_i for each sample. Usually the statistic is the sample mean or the sample median (middle value(s) of the ordered sample). Then the sample mean of the T^*_i is the bagging estimator $\overline{T}^* = \frac{1}{B}\sum_{i=1}^{B} T^*_i$.

The final, Tuesday May 9, 2:45-4:45, covers all 3 exams and all quizzes and homework: 22 sheets of notes. Inference and output for the semiparametric PH, SPH and GCR models is done in the same way (except use PH, SPH or GCR in the appropriate conclusion), but *R* and *SAS* output differ slightly. *R* output uses Z for Wald tests while SAS uses $Z^2 = X^2$ (chisquare for pvalues).

Inference for accelerated failure time models is very similar to that of the PH model, but the *R* and SAS output differ.

So **you need to recognize four types of output for survival regression models**: i) *R* output for semiparametric models, ii) SAS output for semiparametric models, iii) *R* output for accelerated failure time (AFT) parametric regression models, and iv) SAS output for AFT models.