BST 222, Homework 4

Bohao Zou

12/18, 2020

1 Question 1

1.1 (1)

In my thinking, the mother's characteristics that relate with breast feeding are that *mother's alcohol use per month*, *mother's cigarette use*, *mother's age*. The infant's characteristics that relate with breast feeding are that *infant's birth weight*, *number of siblings of infant*.

For the *mother's alcohol use per month*:

- Breast Feed: The medium number of breast feed is 0 and the mean of breast feed is 0.792, the standard deviation of this data is 1.04.
- No breast Feed: The medium number of no breast feed is 0 and the mean of no breast feed is 0.6331. The standard deviation is 1.1124.

From this descriptive analysis we can know that the mother's alcohol use per month is similar in breast feeding or not breast feeding. This may indicate that there is no relationship between those factors.

For the *mother's cigarette use*,

- Breast Feed: The medium number of breast feed is 0 and the mean of breast feed is 0.3794, the standard deviation of this data is 0.6238.
- No breast Feed: The medium number of no breast feed is 0 and the mean of no breast feed is 0.4853. The standard deviation is 0.6948.

From this descriptive analysis we can know that the mother's cigarette using is different between the group of breast feed and no breast feed and the difference is big.

For the *mother's age*,

- Breast Feed: The medium number of breast feed is 22 and the mean of breast feed is 22, the standard deviation of this data is 2.644.
- No breast Feed: The medium number of no breast feed is 21 and the mean of no breast feed is 21.23. The standard deviation is 2.6924.

From this descriptive analysis we can know that the mother's age in years at infant's birth is very similar between those two groups. This may indicate that there is no relationship between those two variables.

For the *infant's birth weight*:

- Breast Feed: The medium number of breast feed is 0 and the mean of breast feed is 0.2392, the standard deviation of this data is 0.4267.
- No breast Feed: The medium number of no breast feed is 0 and the mean of no breast feed is 0.4445. The standard deviation is 0.497.

From this descriptive analysis we can know that the infant's birth weight is different between the group of breast feed and no breast feed but it is not significant. The weight of breast feed is lighter than the no breast feed.

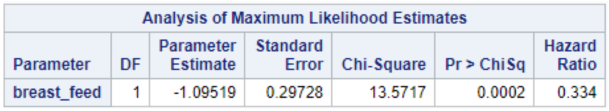
For the *number of siblings of infant*:

- Breast Feed: The medium number of breast feed is 0 and the mean of breast feed is 0.5781, the standard deviation of this data is 0.7985.
- No breast Feed: The medium number of no breast feed is 1 and the mean of no breast feed is 0.7475. The standard deviation is 0.8922.

From this descriptive analysis we can know that number of siblings of infant between the group of breast feed and no breast feed and it is significant different. We can know that there are more siblings of infant in no breast feed group. This may because the milk of mother is not enough to feed for more infants.

1.2 (2)

The result of Cox-PH model to time to hospitalized pneumonia with the indicator of breast feeding as the only covariate in the model is showed below:



Analysis of Maximum Likelihood Estimates

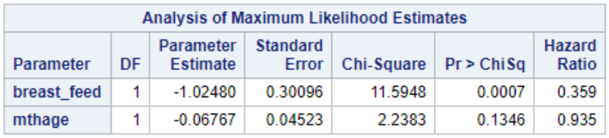
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
breast_feed	1	-1.09519	0.29728	13.5717	0.0002	0.334

Figure 1: The result of Cox-PH model with only breast feeding covariate.

From this table we can know the effect of breast feed is significant at the level of 0.05. The relative risk for breast feed infants compared to no breast feed infants is 0.334 and it is significant. This means that for breast feed infants have probability of infecting pneumonia 0.334 times lower than the probability of infecting pneumonia for no breast feed infants.

1.3 (3)

Adjust for risk factor *mthage*:

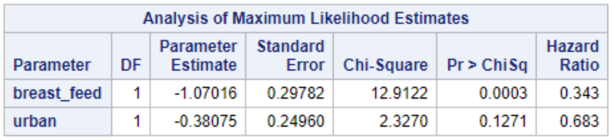


Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
breast_feed	1	-1.02480	0.30096	11.5948	0.0007	0.359
mthage	1	-0.06767	0.04523	2.2383	0.1346	0.935

From this table we can know that from the mother's age increasing, the relative risk of infants infecting pneumonia is decreasing but it is not significant.

Adjust for risk factor *urban*:

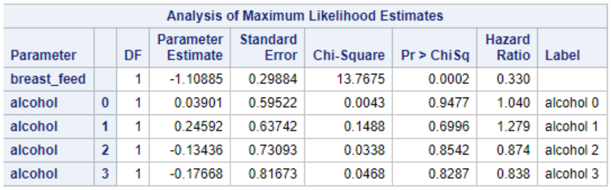


Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
breast_feed	1	-1.07016	0.29782	12.9122	0.0003	0.343
urban	1	-0.38075	0.24960	2.3270	0.1271	0.683

From this table we can know that the relative risk of infecting pneumonia for infants whose mother is in urban compared to those infants whose mother is in rural is 0.683 but it is not significant.

Adjust for risk factor *alcohol*:

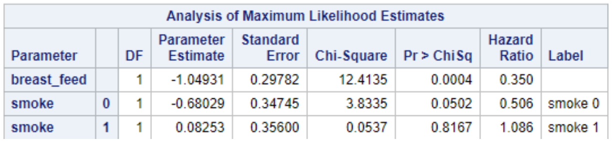


Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
breast_feed		1	-1.10885	0.29884	13.7675	0.0002	0.330	
alcohol	0	1	0.03901	0.59522	0.0043	0.9477	1.040	alcohol 0
alcohol	1	1	0.24592	0.63742	0.1488	0.6996	1.279	alcohol 1
alcohol	2	1	-0.13436	0.73093	0.0338	0.8542	0.874	alcohol 2
alcohol	3	1	-0.17668	0.81673	0.0468	0.8287	0.838	alcohol 3

From this table we can know that the relative risk of infecting pneumonia for infants whose mother does not drink, less than one drink, 1 to 2 drinks and 3 to 4 drinks compared with those infant whose mother drinks more than 4 per month is 1.04, 1.279, 0.874, 0.838 respectively, but it are all not significant.

Adjust for risk factor *smoke*:

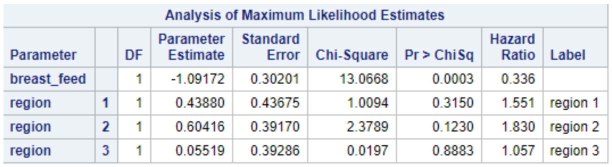


Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
breast_feed		1	-1.04931	0.29782	12.4135	0.0004	0.350	
smoke	0	1	-0.68029	0.34745	3.8335	0.0502	0.506	smoke 0
smoke	1	1	0.08253	0.35600	0.0537	0.8167	1.086	smoke 1

From this table we can know that the relative risk of infecting pneumonia for infants whose mother do not smoke compared to the infants whose mother smoke bigger than 1 pack per day is 0.506 and it is significant under the significant level of 0.1. The relative risk of infecting pneumonia for infants whose mother smoke smaller than 1 pack per day compared to the infants whose mother smoke bigger than 1 pack per day is 1.086 but it is not significant. This indicates that the smoking can increase the risk of infecting pneumonia no matter what you smoke per day.

Adjust for risk factor *Region*:

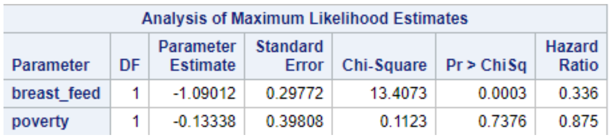


Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
breast_feed		1	-1.09172	0.30201	13.0668	0.0003	0.336	
region	1	1	0.43880	0.43675	1.0094	0.3150	1.551	region 1
region	2	1	0.60416	0.39170	2.3789	0.1230	1.830	region 2
region	3	1	0.05519	0.39286	0.0197	0.8883	1.057	region 3

From this table we can know that the risk factor region has no effect for the children infecting pneumonia under the significant level of 0.1.

Adjust for risk factor *poverty*:

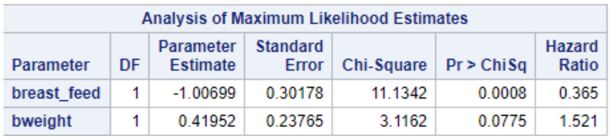


Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
breast_feed	1	-1.09012	0.29772	13.4073	0.0003	0.336
poverty	1	-0.13338	0.39808	0.1123	0.7376	0.875

From this table we can know that the risk factor mother's poverty status have no effect for the children infecting pneumonia under the significant level of 0.1.

Adjust for risk factor *bweight*:

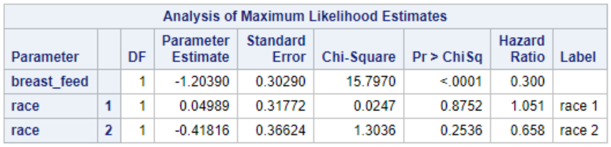


Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
breast_feed	1	-1.00699	0.30178	11.1342	0.0008	0.365
bweight	1	0.41952	0.23765	3.1162	0.0775	1.521

From this table we can know that the relative risk for a infant who birth weight is 1 compared to a birth weight 0 infant is 1.521 and it is significant under the level of 0.1. This means that a heavier infant will have more probability to infect the pneumonia compare with a lighter infant.

Adjust for risk factor *race*:

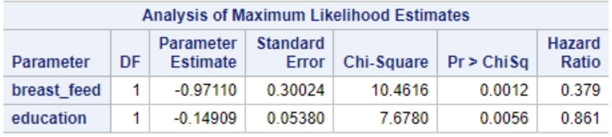


Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
breast_feed		1	-1.20390	0.30290	15.7970	<.0001	0.300	
race	1	1	0.04989	0.31772	0.0247	0.8752	1.051	race 1
race	2	1	-0.41816	0.36624	1.3036	0.2536	0.658	race 2

From this table we can know that the risk factor race of mother has no effect for the children infecting pneumonia under the significant level of 0.1.

Adjust for risk factor *education*:

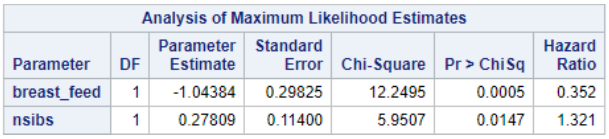


Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
breast_feed	1	-0.97110	0.30024	10.4616	0.0012	0.379
education	1	-0.14909	0.05380	7.6780	0.0056	0.861

From this table we can know that the relative risk for a infant whose mother has 9 years education compared with a infant whose mother has 8 years education is 0.861 and it is significant under the level of 0.1. This means that the infecting risk will decrease when the education years of mother increase.

Adjust for risk factor *nsibs*:



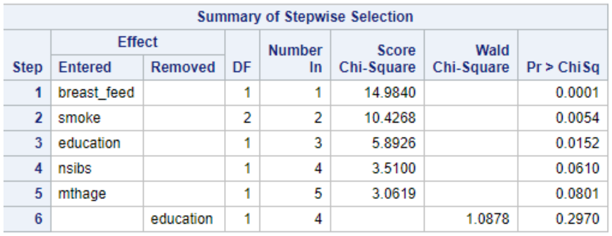
Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
breast_feed	1	-1.04384	0.29825	12.2495	0.0005	0.352
nsibs	1	0.27809	0.11400	5.9507	0.0147	1.321

From this table we can know that the relative risk of a infant who has 5 siblings compared with a infant who has 4 siblings is 1.321 and it is significant under the level of 0.1. This indicates that the risk of infection will increase when the number of siblings increase.

1.4 (4)

In this section, I used the stepwise selection and set the entry P-Value is 0.1 and remove P-Value is 0.1. The reason that why I used stepwise selection is that it is more reasonable than forward and backward selection. The reason that why I set 0.1 as the criteria is that I want more variables to join in the model and can be interpreted. The summary of selection and the final results are showed below:

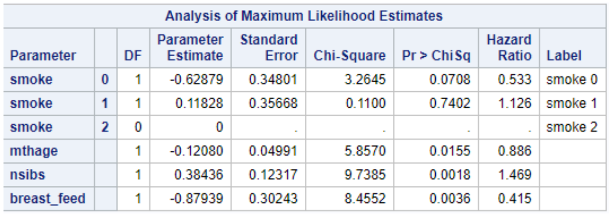


Summary of Stepwise Selection

	Effect						
Step	Entered	Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	breast_feed		1	1	14.9840		0.0001
2	smoke		2	2	10.4268		0.0054
3	education		1	3	5.8926		0.0152
4	nsibs		1	4	3.5100		0.0610
5	mthage		1	5	3.0619		0.0801
6		education	1	4		1.0878	0.2970

Figure 2: Summary of selection procedure.

From the table of summary of selection procedure we can know that the education variable entry into the model at first but removed later. This may indicate that there is a strong multicolinearity in those covarites that selected into model with the education variable.



Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
smoke	0	1	-0.62879	0.34801	3.2645	0.0708	0.533	smoke 0
smoke	1	1	0.11828	0.35668	0.1100	0.7402	1.126	smoke 1
smoke	2	0	0	.	.	.	.	smoke 2
mthage		1	-0.12080	0.04991	5.8570	0.0155	0.886	
nsibs		1	0.38436	0.12317	9.7385	0.0018	1.469	
breast_feed		1	-0.87939	0.30243	8.4552	0.0036	0.415	

Figure 3: The result of final model.

In the end, there are 4 variables in the Cox-PH model and its are all significant under the level of 0.1.

1.5 (5)

From the table of "The result of final model" we can know that the factor *mother doesn't smoke* can decrease the relative risk compared with the *mother smoke more than 1 pack per day* significantly. However, it seems there is no difference between *mother smoke smaller than one pack per day* and *mother smoke more than 1 pack per day.* This means if the mother smoke and no matter the quantity is small or big, it can increase the infection pneumonia risk for their infants.

As the mother's age increasing, the relative risk of infants who may infect pneumonia is decreasing. This may because mother will learn more useful skill or knowledge and will have more sense of responsibility to prevent their infant get any disease.

As the siblings increasing, the relative risk of infecting pneumonia for infants is increasing. This may because their siblings will scatter the energy of their parents and the infant would not get good care.

Finally, the relative risk for the infant who feeds with breast compared with no breast feeding infant is 0.415. This means that breast feeding can decrease the relative risk of infection. This may because the breast feeding protected the infant against hospitalized pneumonia.

2 Question 2

2.1 (1)

I have drew the Kaplan-Meier survival curves and smoothed hazard curves for overall survival by treatment group.

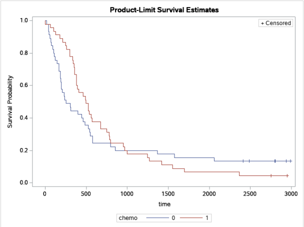


Figure 4: The Kaplan-Meier survival curves

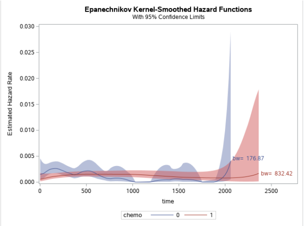
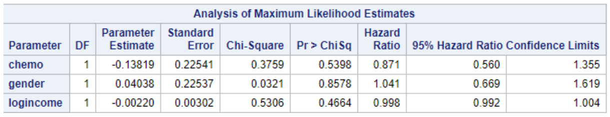


Figure 5: Smoothed hazard curves

From those two plots the estimated survival probability and estimated hazard curves are roughly same between the treatment groups at the same time. This may indicate that the chemotherapy may have the same effect for gastric cancer.

2.2 (2)



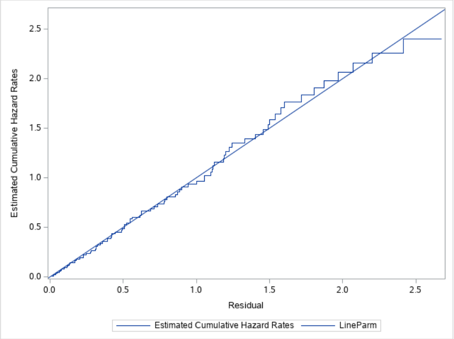
Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
chemo	1	-0.13819	0.22541	0.3759	0.5398	0.871	0.560	1.355
gender	1	0.04038	0.22537	0.0321	0.8578	1.041	0.669	1.619
logincome	1	-0.00220	0.00302	0.5306	0.4664	0.998	0.992	1.004

From this table we can know that the hazard ratio of chemo variable is 0.871 and the associated 95% CI is [0.560, 1.355]. The relative risk for patient who treated with chemo alone compared to the patient who treated with chemotherapy and radiotherapy is 0.871 but it is not significant.

2.3 (3)

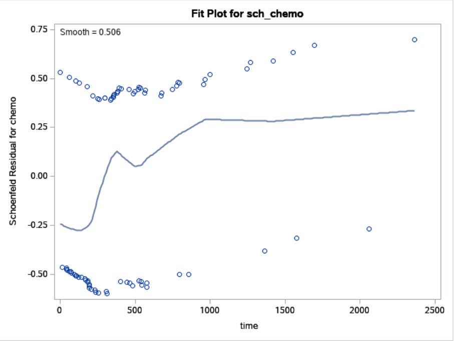
We can use Cox-Snell residual to check the overall fit of the model.



From this plot we can know the model fits the data vary well.

2.4 (4)

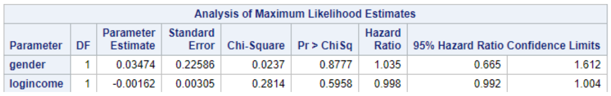
We can use the Schoenfeld residuals against time to evaluate the chemo variable if it fits the PH assumption. A plot shows a non-random pattern against time is evidence of violation of the PH assumption.



From this plot we can know that the plot shows a non-random pattern against time, so the hazard rates for the two treatment groups are not proportianal.

2.5 (5)

We can use stratified Cox PH model to address the issue found in (4).



Analysis of Maximum Likelihood Estimates

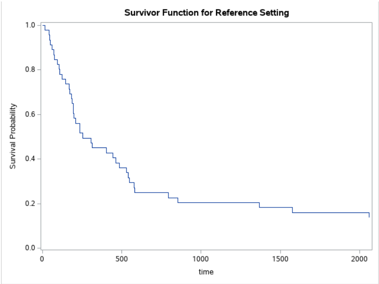
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
gender	1	0.03474	0.22586	0.0237	0.8777	1.035	0.665	1.612
logincome	1	-0.00162	0.00305	0.2814	0.5958	0.998	0.992	1.004

We should check the baseline hazard function $h_{oj}(t)$ where j indicates the stratum to analyze the treatment effects because in the assumption of stratified Cox-PH model, the covariates sharing for all stratum. The partial likelihood for this method is

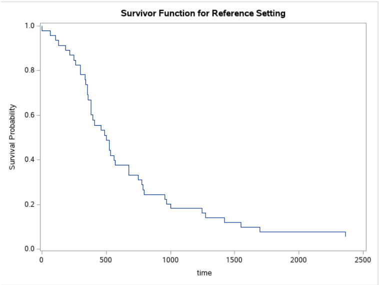
$$LL_j(\overrightarrow{\beta}) = \Sigma_{i=1}^{D}\Sigma_{k=1}^{P}\beta_k Z_{ijk} - \Sigma_{i=0}^{D} log(\Sigma_{m\in R(t_{ij})} exp(\Sigma_{k=1}^{P}\beta_k Z_{mkj}))$$
$$LL(\overrightarrow{\beta}) = \Sigma_{j=1}^{G} LL_j(\overrightarrow{\beta})$$

Where G is the number of stratums and the $LL_j(\overrightarrow{\beta})$ is the log partial likelihood using only the data for those individuals in the $j-th$ stratum. $LL(\overrightarrow{\beta})$ is the partial likelihood for this method.

The baseline survival probability for chem = 0(combination of chemotherapy and radiotherapy) is :



The baseline survival probability for chem = 1(chemotherapy alone) is :



From those two plots, we can know that the survival probability of chemo = 1 is roughly equaled with the survival probability of chemo = 0 under the same time. This may indicate that treated with combination of chemotherapy and radiotherapy and the treated with chemotherapy alone have no effect for survival time for gastric cancer.

2.6 (6)

No, we can't perform the likelihood ratio test to compare Q2 and Q5 models. This is because Q2 model is not a stratified Cox-Ph model but Q5 model is a stratified Cox-PH model. We can't compare them directly. We can change the Q5 model to a normal Cox-Ph model and compare the two models. The H_0 of this compare is $\beta_{chemo} = 0$. The -2 log likelihood of reduced model(Q5, Normal Cox-PH model with only gender and logincome variables) is 614.598. The -2 log likelihood of full model(Q2, Normal Cox-PH model with chemo, gender and logincome variables) is 614.224. $(614.598 - 614.224) = 0.374 \sim \chi^2(1)$. The P-Value bigger than 0.05 and accept the H_0. There is no significant difference between full and reduced model.

2.7 (7)

We can use log likelihood ratio test to examine the assumption of the method used in (5). The assumption is those coefficients are assumed to be the same in each stratum.

The -2 log likelihood of stratified model with gender and logincome variables is $LL(\overrightarrow{\beta}) = 501.710$. The -2 log likelihood of model with gender and logincome variables but only used the chemo=0 data is $LL_0(\overrightarrow{\beta}) = 244.65$. The -2 log likelihood of model with gender and logincome variables but only used the chemo=1 data is $LL_1(\overrightarrow{\beta}) = 256.058$. The likelihood ratio test: $-2(LL(\overrightarrow{\beta}) - LL_0(\overrightarrow{\beta}) - LL_1(\overrightarrow{\beta})) = (501.710 - 244.65 - 256.058) = 1.012 \sim \chi^2(2)$. The P-Value is 0.602902. It is not significant under 0.05 level. So, the assumption holds.

3 Appendix

```
proc import datafile = "/folders/myshortcuts/MyFolder/KM1.13.csv" out=KM;

* cox reg;
proc phreg data=KM;
model chldage*hospital(0) = breast_feed;
run;

proc phreg data=KM;
class alcohol;
class smoke;
class region;
class race;
model chldage*hospital(0) = breast_feed mthage;
run;

proc phreg data=KM;
class alcohol;
class smoke;
class region;
class race;
model chldage*hospital(0) = breast_feed urban;
run;


proc phreg data=KM;
class alcohol;
class smoke;
class region;
class race;
model chldage*hospital(0) = breast_feed alcohol;
run;

proc phreg data=KM;
class alcohol;
class smoke;
class region;
class race;
model chldage*hospital(0) = breast_feed smoke;
run;

proc phreg data=KM;
class alcohol;
class smoke;
class region;
class race;
model chldage*hospital(0) = breast_feed region;
run;

proc phreg data=KM;
class alcohol;
class smoke;
class region;
class race;
model chldage*hospital(0) = breast_feed poverty;
run;

proc phreg data=KM;
class alcohol;
class smoke;
class region;
class race;
model chldage*hospital(0) = breast_feed bweight;
run;


```

```

proc phreg data=KM;
class alcohol;
class smoke;
class region;
class race;
model chldage*hospital(0) = breast_feed race;
run;

proc phreg data=KM;
class alcohol;
class smoke;
class region;
class race;
model chldage*hospital(0) = breast_feed education;
run;


proc phreg data=KM;
class alcohol;
class smoke;
class region;
class race;
model chldage*hospital(0) = breast_feed nsibs;
run;


ods noproctitle;
ods graphics / imagemap=on;

proc phreg data=WORK.KM;
  class alcohol smoke region race / param=glm;
  model chldage*hospital(0)=alcohol smoke region race mthage poverty urban
    education bweight nsibs breast_feed / selection= stepwise slentry=0.1  slstay=0.1;
run;

*question 2;
proc import datafile = "/folders/myshortcuts/MyFolder/GASTRIC.csv" out=gas;

* smoothed hazard function;
proc lifetest data=gas method=KM plots=hazard(cl);
time time*status(0);
strata chemo;
run;

proc phreg data=gas;
  model time*status(0)=chemo gender logincome / rl;
run;


* Cox-snell residuals plot for original data;
proc phreg data = gas;
model time*status(0)=chemo gender logincome / rl;
output out=plot1_1 logsurv=logsurv1 /method = ch;

data plot1_1;
set plot1_1;
snell = -logsurv1;
cons = 1;

proc phreg data=plot1_1;
model snell*status(0) = cons;
output out = plot1_2 logsurv= logsurv2/method=ch;

data plot1_2;
set plot1_2;
cumhaz = - logsurv2;

```

```
proc sort data=plot1_2;
by snell;

proc sgplot data= plot1_2;
step y=cumhaz x=snell /MARKERFILLATTRS=(color="red");
lineparm x=0 y=0 slope=1; /** intercept, slope **/
label cumhaz = "Estimated Cumulative Hazard Rates";
label snell = "Residual";
run;

* Schoenfeld residual;
proc phreg data = gas;
model time*status(0)=chemo gender logincome / rl;
output out=schoen ressch=sch_chemo;

proc loess data= schoen plots=fitplot;
model sch_chemo=time;
run;

*stratified Cox ph model;
proc phreg data=gas;
model time*status(0) = gender logincome / rl;
strata chemo;
run;


* not a stratified cox ph model;
proc phreg data=gas;
model time*status(0) = gender logincome / rl;
run;


*stratified Cox ph model;
proc phreg data=gas;
model time*status(0) = gender logincome / rl;
strata chemo;
run;
*stratified Cox ph model;
proc phreg data=gas;
where chemo = 0;
model time*status(0) = gender logincome / rl;
run;
*stratified Cox ph model;
proc phreg data=gas;
where chemo = 1;
model time*status(0) = gender logincome / rl;
run;

proc phreg data = gas plots=survival;
where chemo = 0;
model time*status(0) = gender logincome;
baseline;
run;

proc phreg data = gas plots=survival;
where chemo = 1;
model time*status(0) = gender logincome;
baseline;
run;
```