University of Toronto Scarborough
Department of Computer and Mathematical Sciences
STAD29 / STA 1007 (K. Butler), Midterm Exam
February 13, 2016

Aids allowed:

- My lecture overheads (slides)
- The R "book"
- Any notes that you have taken in this course
- Your marked assignments
- Non-programmable, non-communicating calculator

Before you begin, complete the signature sheet, but sign it only when the invigilator collects it. The signature sheet shows that you were present at the exam.

This exam has 13 numbered pages of questions. Please check to see that you have all the pages.

In addition, you should have an additional booklet of output to refer to during the exam. Contact an invigilator if you do not have this.

Answer each question in the space provided (under the question). If you need more space, use the backs of the pages, but be sure to draw the marker's attention to where the rest of the answer may be found.

The maximum marks available for each part of each question are shown next to the question part. In addition, the total marks available for each page are shown at the bottom of the page, and in the table on the next page.

The University of Toronto's Code of Behaviour on Academic Matters applies to all University of Toronto Scarborough students. The Code prohibits all forms of academic dishonesty including, but not limited to, cheating, plagiarism, and the use of unauthorized aids. Students violating the Code may be subject to penalties up to and including suspension or expulsion from the University.

Last name: _____

First name: _____

Student number: _____

For marker's use only:

Page	Points	Score
1	8	
2	8	
4	7	
6	6	
7	6	
8	11	
10	12	
12	5	
13	9	
Total:	72	

1. The North American professional basketball league, called the NBA, had 30 teams in the 2008–2009 season. What factors affect the average number of fans that come to each team's games? (This is called the team's "attendance".) Three possible factors are:

   - the recent success of the team, measured as number of wins in the 82-game 2008–2009 season.
   - the size of the city. This was measured as the size of the team's TV market (as a percentage of all households in the US). This variable is called "market share".
   - Whether or not the team has played in the NBA finals in the last five years. (This is another measure of recent success).

   The data are shown in Figure 1 in the booklet of code and output.

   (a) (2 marks) A regression model is shown in Figure 2. Would you say that the factors named above together give a good, moderate or poor prediction of team attendances? Explain briefly.

   My answer: I would look at the R-squared. Here, that is about 30%, which is not great, but not worthless either. Alternatively, or additionally, you can look at the global F-test, which has a P-value of 0.022, which is smaller than 0.05, so that *something* is helping to predict attendance. I would use the word "moderate" to describe this fit, but if you want to call it "poor", that's fine too, as long as you express at least one reason for doing so.

   (b) (2 marks) Interpret the slope of `wins`: that is, express in words in the context of the data what it is telling you.

   My answer: Each additional win adds an average of 68 fans to a team's attendance, other things being equal.

   (c) (3 marks) Why is the slope of `finals` shown in the output as `finalsyes`? Interpret this value.

   My answer: `finals` is a categorical variable taking the values `yes` and `no`. When it is added to a regression, the alphabetically first level, `no`, is used as a baseline and the other level(s), here `yes`, are expressed relative to that. Here, the slope of `finalsyes` is 1038, which means that a team having played in the NBA finals in the last five years adds 1038 to the team's average attendance, other things being equal.

   (d) (1 mark) Figure 3 shows the residual plots from the regression in Figure 2. What was the purpose of the `par(mfrow=c(2,2))` above the plots?

   My answer: To display the plots in a 2×2 grid rather than one after the other.

(e) (2 marks) What do you conclude from the top left plot in Figure 3? Does that indicate any problems with the regression? Explain briefly.

My answer: This plot is of the residuals against the fitted values. This should be a random mess of points, if the straight line model is appropriate. The lowess curve is showing a hint of an up-and-down curved pattern, but my immediate impression is that the points on the plot are randomly scattered, so I think that the regression we fit is appropriate. Having said that, the plot doesn't really have any points in the top right: for a fitted attendance around 20,000, the residuals are mainly close to zero or negative. So a variety of answers is possible here.

As for the impact on the appropriateness of the regression: if you found a problem, the regression is not appropriate as it stands and it needs to be fixed up; if you found no problems, all is good.

(f) (2 marks) What do you conclude from the top right plot in Figure 3? Does that indicate any problems with the regression?

My answer: The residuals are very close to being normally distributed, which is an assumption that we are making for the regression. So the regression is appropriate.

You need to address both parts to get both marks.

Alternatively, there is the smallest hint that the most negative residuals are not negative enough. (This also shows up on the top left plot: the most positive residuals are around 4000 in size, and the most negative residuals are only around -3000.)

(g) (2 marks) What do you conclude from the bottom left plot in Figure 3?

My answer: Again, a couple of ways to go:

The lowess curve is going slightly downhill, so there is the tiniest evidence of fanning-in: residuals are getting (slightly) smaller in size as the fitted values get bigger.

The lowess curve is basically flat and the points are scattered all over the plot, so there is no evidence that the size of the residuals changes as the fitted values change.

I like the second one better, but I'll take the first if you have all the pieces together: what there is, and what it tells us.

My opinion is that the fanning-in, such as it is, is connected with the absence of points top right of the first residual plot. There are really no teams with large average attendances whose actual attendance is larger than predicted. This might be connected with the capacity of the buildings where these teams play: the building is usually sold out, so the attendance *can't* get any bigger. I did an analysis a long time ago of soccer attendances in England, which said the same thing, that some teams could have gotten substantially bigger attendances if their stadiums had been bigger.

(h) (2 marks) Make a recommendation for the next thing to do in this analysis, if anything. Support your recommendation briefly.

My answer: Find something and support it. There are several possibilities, including these:

- If you found something wrong in the residual plots, suggest a fix. For example, if you thought the top-left plot indicated a curve, then suggest fitting a curve, eg. by adding something squared, or transforming the response (attendance), for example by running Box-Cox.
- Eliminate a non-significant variable in the regression. I would try taking out `finals`, since its P-value is the biggest. (The variable is called `finals`, and the slope of it that

was fit is called `finalsyes`.)

- Saying "we are done" is also fine, if you can support that (which I think you will have trouble doing).

As ever, anything sensible is rated according to its sensibleness.

2. Some scientists were studying the ozone layer over the Antarctic. These scientists developed a measure of the degree to which oceanic phytoplankton production (labelled `inhibit` in the data) is inhibited by exposure to ultraviolet radiation (`uvb` in the data set). The statistical question is how, if at all, `inhibit` is related to `uvb`. To complicate matters, measurements were taken at two different depths, labelled `surface` and `deep`, and the scientists wanted to see whether depth had any impact as well. The first task the scientists wanted to do is to make a plot of `inhibit` against `uvb`, with the points labelled according to depth.

   Unfortunately, the data come to us in the form shown in Figure 4. There are two columns of `inhibit`, one for depth `deep` and one for depth `surface`. Likewise, there are two columns of `uvb` measurements.

   Underneath the listing of the data, I load `dplyr` in such a way as not to show us those messages it would otherwise show us.

   (a) (3 marks) There are four lines of `dplyr` code in Figure 5. (There are actually only two lines of R code altogether, but each one is split into two to fit on the paper). Describe in words what the first two printed lines, from `ozone %>%` to `-> deep` inclusive, are doing. Do this in such a way that someone who does not understand `dplyr` can see what is being done.

   My answer: Take the data frame `ozone`, and from it create two new variables `inhibit` and `uvb` which are the values of those variables for the deep-level measurements. Then select only those two variables that we just created (getting rid of the others), and save them in a new data frame called `deep`.

   Your explanation ought not to use the word "mutate", since, unless you know about `dplyr`, you won't know what that means in this context.

   The other two lines do the same thing, but for the surface-level measurements. So we now have two data frames, `deep` and `surface` that each contain two columns, `inhibit` and `uvb`, for the measurements taken at their respective depths.

   (b) (2 marks) Figure 6 contains some code and output that uses the data frames `deep` and `surface` created in Figure 5. I do not expect you to have seen `bind_rows` before, but try to explain in words what it is doing. There are two specific things that I would like you to observe.

   My answer: The name `bind_rows` is a bit of a clue: it takes the rows of `deep` and the rows of `surface` and puts them one above the other. That's the first thing, and it's the same thing that `rbind` would do (`rbind` is like `cbind` but stacks the data frames on top of each other rather than side by side).

   The second thing, which is the extra thing that `bind_rows` does, is to create a new column, here called `depth`, that says which data frame each new row came from. The first ten came from `deep` and the rest from `surface`. I had to figure out how to do this. The key is that `.id="depth"` on the end: this says "create a new column called `depth` that says which data frame each row came from". I wanted to get their names in there, so I had to do that `deep=deep` thing; this says to create a *name* `deep` (the first `deep`) and get its values from the data frame `deep` (the second one). Otherwise, `depth` would have contained the numbers 1 and 2, not the names. This is the same idea as when you create a data frame using `data.frame` and you want to make sure that the columns have the right names.

   (c) (2 marks) Figure 7 shows some code to make a `ggplot` plot, and the output that it produces. (The warnings below the code are about the missing values in the data frame `v`. You do not need to worry about these.) Explain briefly what role the `aes` plays here.

   My answer: This is the "what to plot". Specifically, it says what to put on the x-axis of whatever plot will be produced (`uvb`), what goes on the y-axis (`uvb`) and that we will want to

distinguish by colour (somehow) the different `depth` values.
How the different depth values will be distinguished by colour is left up to the plot that is produced. But the `aes` is making sure that it happens somehow.

(d) (2 marks) In Figure 7, what role does the `geom_smooth` play here?

My answer: This produces a line or curve through the points. The default is a lowess curve, but here, since I put in the `method="lm"`, we get regression lines instead. (We happen to get one for each group because of the `colour` in the `aes`, but that's not central to what the `geom_smooth` does.)

(e) (1 mark) Suppose I had wanted to make a scatterplot of all the points, with *one* regression line for all the data. How would I modify the code in Figure 7 to achieve that?

My answer: Take out the `colour=depth` from the `aes`. That's the only change we need to make.

(f) (3 marks) What do you conclude about the relationship, if any, between `uvb` and `inhibit`? Is the nature of the relationship the same or different for the two different depths?

My answer: For both groups, a larger `uvb` is associated with a larger `inhibit` (the points are reasonably close to the lines). But `inhibit` increases faster with `uvb` for the `deep` measurements than the `surface` ones (the line is steeper). If the lines were parallel, there would be a consistent relationship between the two groups, but they are not.

3. Back in the 1800s, people would try to bring their favourite bird species from their homeland to where they were living. They would release the birds, and hope that the birds would become successfully established in the new country. This was done a great deal in New Zealand. In 1996, an exploratory study was done, as an attempt to determine what factors might lead to a bird species becoming successfully established. They looked at 79 bird species that had been introduced to New Zealand prior to 1907, and measured a number of variables related to the species and introduction(s), as follows:

   - `Species`: abbreviated name of species
   - `Status`: whether species still present (1) or absent (0) in New Zealand (response variable)
   - `Length`: female body length (mm)
   - `Mass`: female body mass (g)
   - `Range`: geographic range (% of area of Australia. Why not New Zealand I have no idea.)
   - `Migr`: migratory (3), somewhat migratory (2) or not migratory (1). ("Migratory" means the bird flies to a warmer climate in winter or a cooler one in summer.)
   - `Insect`: number of months of year with insects in diet
   - `Diet`: eats plants (1), everything (2), meat (3)
   - `Clutch`: number of eggs per breeding episode
   - `Broods`: number of breeding episodes per year
   - `Wood`: uses woodland areas frequently (1), infrequently (0)
   - `Upland`: uses upland areas (hills or mountains) frequently (1), infrequently (0)
   - `Water`: uses water frequently (1) or infrequently (0)
   - `Release`: number of times birds of the species released
   - `Indiv`: total number of individual birds released.

   I have treated the variables `Migr` and `Diet` as numeric. Although they are really categorical, we can think of an implied underlying continuous scale in each case. (For example, a bird species might have a greater or lesser tendency to eat meat.)

   The structure of the data is shown in Figure 8. The original data had some missing values. I deleted all the observations with any missing values.

   (a) (2 marks) Why is logistic regression a plausible technique to use here?

   My answer: The response variable `Status` is categorical: it only takes the values 1 (present) or 0 (absent). With a categorical response variable, logistic regression is plausible (and regular regression is not).

   (b) (2 marks) In a logistic regression, what probability will R model? Explain briefly.

   My answer: Here, the reponse variable is a factor (or something behaving like a factor), so the probability of the *second* alphabetical level of the factor will be modelled. Here, that is whether the species is present (1).

   (c) (2 marks) A logistic regression is fit in Figure 9. Why do you think I chose to begin with this model? Explain briefly.

   My answer: I started with *all* the explanatory variables, in the hopes of finding some of them that make a difference. (I didn't know which ones were likely to be relevant.) I can then eliminate the ones that seem to have nothing to say (but that's the next part).

(d) (3 marks) A second analysis is shown in Figure 10. Why do you think I chose to do this analysis next? Bearing in mind the exploratory nature of this problem (I am trying to find some potentially interesting factors in determining whether an introduced bird species becomes successfully established or not), do you think I have now found a good model, or not? Explain briefly.

My answer: Check the variables I removed: they were all the ones with largest P-values, and therefore the ones that had least to add to the logistic regression.

I think I have now found a good model, because the P-values I have are now all fairly small (which they weren't before). I'm thinking that using $\alpha = 0.10$ is sensible, because I'm doing an exploratory analysis ("potentially interesting"), though it is also at least somewhat reasonable to insist on the 0.05 cutoff and think about removing `Wood` now. I wouldn't remove more than one explanatory variable, because there are several P-values close to 0.05 that might become significant once `Wood` is removed. (Taking out several variables to my mind is too dangerous at this point, though you could argue for taking several out and then testing using `anova` to see whether it was a good idea.)

(e) (2 marks) Look at the test shown in Figure 11. What do you conclude from it, in the context of the data?

My answer: This test is comparing the two logistic regression models shown in Figures 9 and 10. The null hypothesis is that the two models are equally good, with the alternative being that the bigger model in Figure 9 is better. In this case the null is not rejected, so the two models are equally good and we can go with the smaller, simpler one. That is, it is reasonable to remove all the variables I did, namely `Length, Range, Insect, Clutch, Broods, Water, Release`.

(f) (2 marks) In Figure 10, I used `update`. Explain briefly, in words, what that is doing here.

My answer: It is an alternative way of specifying a model: rather than writing out a whole model, we can write the *change* from a previous model. In this case, it says "take the model in Figure 9, and from it, remove the seven variables listed, `Length, Range, Insect, Clutch, Broods, Water, Release`." (The response variable and the kind of analysis are left unchanged.)

(g) (2 marks) Look at the output in Figure 10. Does this indicate that a species for which more individuals were released is more, less or equally likely to be still present in New Zealand, other things being equal? How can you tell? Explain briefly.

My answer: This is talking about the effect of the explanatory variable `Indiv`. This is significant (small P-value) and positive. So when more individuals were released, the response event is more likely to happen. The response event is that the species is still present in New Zealand.

This is entirely what you'd expect: if a large number of individuals was released, enough of them are likely to survive to make the species viable. If only a small number of individuals was released, they might not make it.

(h) (2 marks) In Figure 10, the slope coefficient for `Upland` is significantly negative. What does that mean, in the context of the data?

My answer: When `Upland` is larger, the probability of the species still being present is smaller, all else being equal. What does `Upland` being larger mean? Well, according to the definition of the variable, this is 1 if the species uses hills or mountains frequently and 0 if infrequently. So,

a species that uses hills and mountains frequently is less likely to still be present, all else being equal.

In fact, the slope coefficient is large in size (-5.122). This means that our best guess is that a species that uses upland areas frequently (as opposed to one that uses upland areas infrequently) has its odds of still being present multiplied by

```
exp(-5.122)
```

```
## [1] 0.005964083
```

or about 1 in 200. This is a very dramatic change. (The slope coefficient has a large standard error, so it is imprecisely estimated, which means that the actual effect might be quite a bit different from this.)

Now that I think about it, this would be a great opportunity to use `step` (as from one of the practice questions on Assignment 3), since we have a lot of potential explanatory variables and we want to get rid of the worthless ones. Here's how it would go:

```
nzbirds=read.table("nzbird.txt",header=T)
nzbirds.1=glm(Status~Length+Mass+Range+Migr+Insect+Diet+
  Clutch+Broods+Wood+Upland+Water+Release+Indiv,
  data=nzbirds,family="binomial")
step(nzbirds.1,trace=0)
```

```
##
## Call:  glm(formula = Status ~ Mass + Migr + Insect + Wood + Upland +
##     Indiv, family = "binomial", data = nzbirds)
##
## Coefficients:
## (Intercept)         Mass         Migr       Insect         Wood
##   -3.437608     0.001939    -2.023985     0.270468     1.948944
##      Upland        Indiv
##   -4.730656     0.013812
##
## Degrees of Freedom: 66 Total (i.e. Null);  60 Residual
## Null Deviance:     90.34
## Residual Deviance: 28.03  AIC: 42.03
```

This is very like (but not identical to) my model `nzbirds.2`: it has `Insect` instead of `Diet`, but is otherwise the same. It is a bit more scientific than my model, which was obtained by throwing out a bunch of apparently worthless explanatory variables and hoping that I hadn't thrown out too much. The slopes of the variables the two models have in common are more or less the same.

4. On January 28, 1986, the space shuttle Challenger took off and exploded, killing all the astronauts aboard. Might it have been possible to anticipate the problems and stop the launch on that day? Data were collected on all the space shuttle flights, as follows:

   - `flight`: flight number, 1 through 25 (two flights had missing data and were omitted from the data set)
   - `distress` (response): The number of "thermal distress incidents" in which hot gas damaged the joint seals of a flight's booster rockets. Damage to the joint seals helped lead to the Challenger disaster. Categorized as 0, 1–2 or 3+ (3 or more): the higher the distress, the riskier the flight.
   - `temp`: Calculated joint temperature at launch time (degrees F).
   - `date`: the date (number of days since Jan 1, 1960). There may have been changes in the shuttle program or the hardware over time that may have had an impact on `distress`.
   - `z.computed.`: ignore.

   A summary of the data is shown in Figure 12.

   (a) (2 marks) What is it about the data that makes the function `polr` more suitable for carrying out the analysis than `multinom`?

   My answer: The response variable `distress` is ordered: `0` is less than `1--2` which is less than `3+`. With an ordered response, we use `polr` (with an unordered nominal response we would use `multinom`).

   (b) (2 marks) In Figure 13, some models are fitted. What do you conclude, in the context of the data, from Figure 14?

   My answer: This is comparing a model with both the explanatory variables to one with date only. So it's testing the effect of temperature. Since the P-value is small, the bigger model is needed: temperature should stay in the model.

   (c) (2 marks) What do you conclude, in the context of the data, from Figure 15?

   My answer: This is comparing both explanatory variables to temperature only. This P-value is also small, so date should stay in the model as well.

   To summarize both the last two parts, we need both explanatory variables in the model.

   (d) (4 marks) Figure 16 shows some predictions for representative values of `temp` and `date`. (These values are approximately the first and third quartiles of each variable.) Describe the effects of (i) increased temperature and (ii) increased date on the response variable.

   My answer: Hold one variable fixed and look at the effect of increasing the other on the three probabilities.

   For increasing `temp`, the effect at either `date` is that the probability of a distress of `3+` is sharply decreased and the other two probabilities go up. That is to say, the predicted *amount* of distress goes down. (Things are better at a higher temperature.)

   For increasing `date`, the effect at either `temp` is of a large increase in the probability of `3+` and a decrease in the other two probabilities. That is, the predicted amount of distress goes *up*. (Things evidently were getting old, which was the reason for including the date in the analysis.)

   (e) (2 marks) On the date of the Challenger launch, which was day 9524 on this scale, the temperature was 31 F. A prediction is shown in Figure 17. (Note that the format is different for some reason:

Question 4 continues…

This page: _____ of possible 12 points.

the predictions go down the page instead of across.) What would you say to the people in charge of the launch, and why?

My answer: Looking at the predictions, there is a 99.96% chance of being 3 or more distress incidents (and a correspondingly tiny chance of being 2 or fewer distress incidents). This means that the flight would be very risky, according to the question preamble. I think we are entitled to tell the people in charge of the launch that launching now is too risky and it would be better to wait (and try the launch again later when it is warmer).

The physicist Richard Feynman demonstrated at a press conference that the key component of the joint seals, something called an O-ring, would become brittle and inflexible at low temperatures. He did this by dropping an O-ring into his glass of ice water!

From what I remember hearing about this, there was a lot of political pressure on NASA to get Challenger launched, and this pressure overcame the scientific consensus, which was that it was too cold and the launch should be delayed. The launch had already been delayed several times. (31 degrees was by far the coldest `temp` value for all the space shuttle launches, so, for us, there is the issue of extrapolation as well.)

This is rather poignant: `http://www.npr.org/sections/thetwo-way/2016/01/28/464744781/30-years-after-disaster-challenger-engineer-still-blames-himself`.

5. Primary biliary cirrhosis is a disease in which the bile ducts in the liver are slowly destroyed. If this happens, harmful substances can build up in the liver and lead to scarring of the liver tissue (which is what "cirrhosis" is). A study was carried out at the Mayo Clinic between 1974 and 1984. This was a placebo-controlled trial of the then-new drug D-penicillamine (with some patients receiving a placebo). In the study, a number of other variables were measured that might have some impact on a subject's survival. (I know nothing about what these other variables are.) The aim of the study was to see whether patients who took the new drug tended to survive longer than patients on the placebo, after adjusting for the effects of the other variables. The structure of the data is shown in Figure 18. The variables are these:

   - `id`: subject ID
   - `days`: survival time (days) after diagnosis
   - `status`: what happened to the patient: alive at last doctor visit (0), removed from study to have liver transplant (1), dead (2).
   - `drug`: D-penacillamine (1), placebo (2).
   - `age`: patient's age at diagnosis (days).
   - `edema`: yes (1), no (0). There is also a value 0.5, "edema resolved by diuretics". We treat this variable as a continuous numeric variable.
   - `bilirubi`: serum bilirubin (mg/dl).
   - `albumin`: serum albumin (mg/dl).
   - `prothom`: prothrombin time (seconds).

   (a) (3 marks) Look at Figure 19. What is happening there? Your answer should address three things: (i) why I am going to need this variable `y`, (ii) the purpose of the `status==2`, (iii) precisely why some of the values of `y` have `+` next to them.

   My answer: (i) I am going to do a survival analysis, which needs a response variable, and `y` will be that. (ii) `Surv` needs two things: the variable representing the survival time, here `days`, and the condition representing the event of interest, here death, which occurred when `status` is 2. (iii) The values with `+` next to them are "censored", that is to say, that subject did not die: either they were still alive at the end of the study, or they left the study to have a liver transplant.

   The nice thing about the proportional-hazards model is that as long as the reason for censoring has nothing to do with the event of interest (for example, if the sickest subjects were the ones being taken off to have a liver transplant, that would be a problem), then the reason for censoring is irrelevant. We are assuming that being chosen for a liver transplant is not related to the likelihood of death.

   (b) (2 marks) Are the researchers looking for a coefficient of `drug` that is positive or negative? Explain briefly.

   My answer: For the data, the new drug is labelled 1 and the placebo is labelled 2. Increasing `drug` ought to make the hazard of death *higher*, because the new drug ought to be better at delaying death than the placebo. So the coefficient ought to be *positive*.

Question 5 continues…

This page: _____ of possible 5 points.

(c) (2 marks) A Cox proportional hazards model is fitted in Figure 20. What does this output tell you about any effect of `drug`? Explain briefly.

My answer: It says that, after adjusting for the effects of all the other variables, there is *no* effect of drug, since the P-value is far from small at 0.947. In fact, the slope coefficient is a little *negative*, so the placebo is actually slightly *better* at delaying death, but the effect is very small.

(d) (2 marks) *Not* including `drug`, do any or all of the other variables have a significant effect on survival time? Explain briefly.

My answer: They all do. All of them have P-values less than 0.05, in some cases a *lot* less than 0.05.

(e) (3 marks) For *each* of the significant variables in the model, would a high value or a low value be associated with surviving for a long time?

My answer: First off, we need to ignore `drug`, since that has no relation with survival.

The model is actually predicting *hazard of death*, so a high value is bad and a low value is good when it comes to death. That means that for variables with positive slope coefficients we want a *low* value and for variables with negative slope coefficients we want a *high* value.

Thus, these things are associated with surviving for a longer time:

- low `age`
- low `edema`
- low `bilirubi`
- high `albumin`
- low `prothom`

We normally assess this kind of thing by looking at a graph, but I figured I'd give you a change.

(f) (2 marks) Figure 21 shows the set-up for some predictions of survival curves, using representative values for the variables, and Figure 22 shows the plot of those predicted survival curves. Explain how the plot is consistent with *two* things that you have previously seen in this question.

My answer: Survival curves up and to the right are associated with better survival. The red and blue survival curves are better than the green and black ones. The red and blue curves are associated with the lower age, so being of lower age is associated with better survival. This is what we saw in part (e) (the first thing I said). Also shown on the plot is the effect of `drug`: at age 15000 days, comparing the red and blue curves shows the near-zero effect of drug, since the survival curves are more or less coincident. Likewise the green and black curves for age 20000 days.

In short, a significant age effect (lower age better) and a non-existent drug effect.

Final observations: What I called "dead" is rather charmingly described in the textbook from which I got these data as "event (nonsurvival)". Also, this particular drug features nowhere that I could find in current descriptions of treatment of primary biliary cirrhosis. That suggests to me that this study was important in eliminating this apparently promising but actually worthless drug from consideration for treating this disease.