432 Class 25 Slides

github.com/THOMASELOVE/432-2018

2018-04-17

Preliminaries

```
library(skimr)
library(rms)
library(survival)
library(OIsurv)
library(survminer)
library(broom)
library(tidyverse)
```

Today's Agenda

- Regression on Time-to-event data
  - Cox Proportional Hazards Model
- Some Loose Ends

Survival Analysis / Cox Regression

A Survival Analysis Example

Source: Chen and Peace (2011) *Clinical Trial Data Analysis Using R*, CRC Press, section 5.1

```
brca <- read.csv("data/breast_cancer.csv") %>% tbl_df
```

The brca trial

The brca data describes a parallel randomized trial of three treats, adjuvant to surgery in the treat of patients with stage-2 carcinoma of the breast. The three treat groups are:

- S+CT = Surgery plus one year of chemotherapy
- S+IT = Surgery plus one year of immunotherapy
- S+CT+IT = Surgery plus one year of chemotherapy and immunotherapy

The measure of efficacy were "time to death" in weeks. In addition to treat, our variables are:

- trial_weeks: time in the study, in weeks, to death or censoring
- last_alive: 1 if alive at last follow-up (and thus censored), 0 if dead
- age: age in years at the start of the trial

brca tibble

```
# A tibble: 31 x 5
   subject treat   trial_weeks last_alive   age
   <fct>   <fct>         <int>      <int> <int>
 1 A01     S+CT            102          0    55
 2 A02     S+IT            192          0    62
 3 A03     S+CT+IT          73          0    72
 4 A04     S+CT             58          1    48
 5 A05     S+CT             48          1    26
 6 A06     S+IT            182          1    52
 7 A07     S+IT            196          1    50
 8 A08     S+CT            177          1    49
 9 A09     S+IT            191          1    62
10 A10     S+CT+IT          36          0    60
# ... with 21 more rows
```

Analytic Objectives

This is a typical right-censored survival data set with interest in the comparative analysis of the three treats.

1. Does immunotherapy added to surgery plus chemotherapy improve survival? (Comparing S+CT+IT to S+CT)
2. Does chemotherapy add efficacy to surgery plus immunotherapy? (S+CT+IT vs. S+IT)
3. What is the effect of age on survival?

Create survival object

- `trial_weeks`: time in the study, in weeks, to death or censoring
- `last_alive`: 1 if alive at last follow-up (and thus censored), 0 if dead

So `last_alive` = 0 if the event (death) occurs.

What's next?

Create survival object

- `trial_weeks`: time in the study, in weeks, to death or censoring
- `last_alive`: 1 if alive at last follow-up (and thus censored), 0 if dead

So `last_alive` = 0 if the event (death) occurs.

```
brca$S <- with(brca, Surv(trial_weeks, last_alive == 0))

head(brca$S)
```

```
[1] 102  192   73   58+  48+ 182+
```

Build Kaplan-Meier Estimator

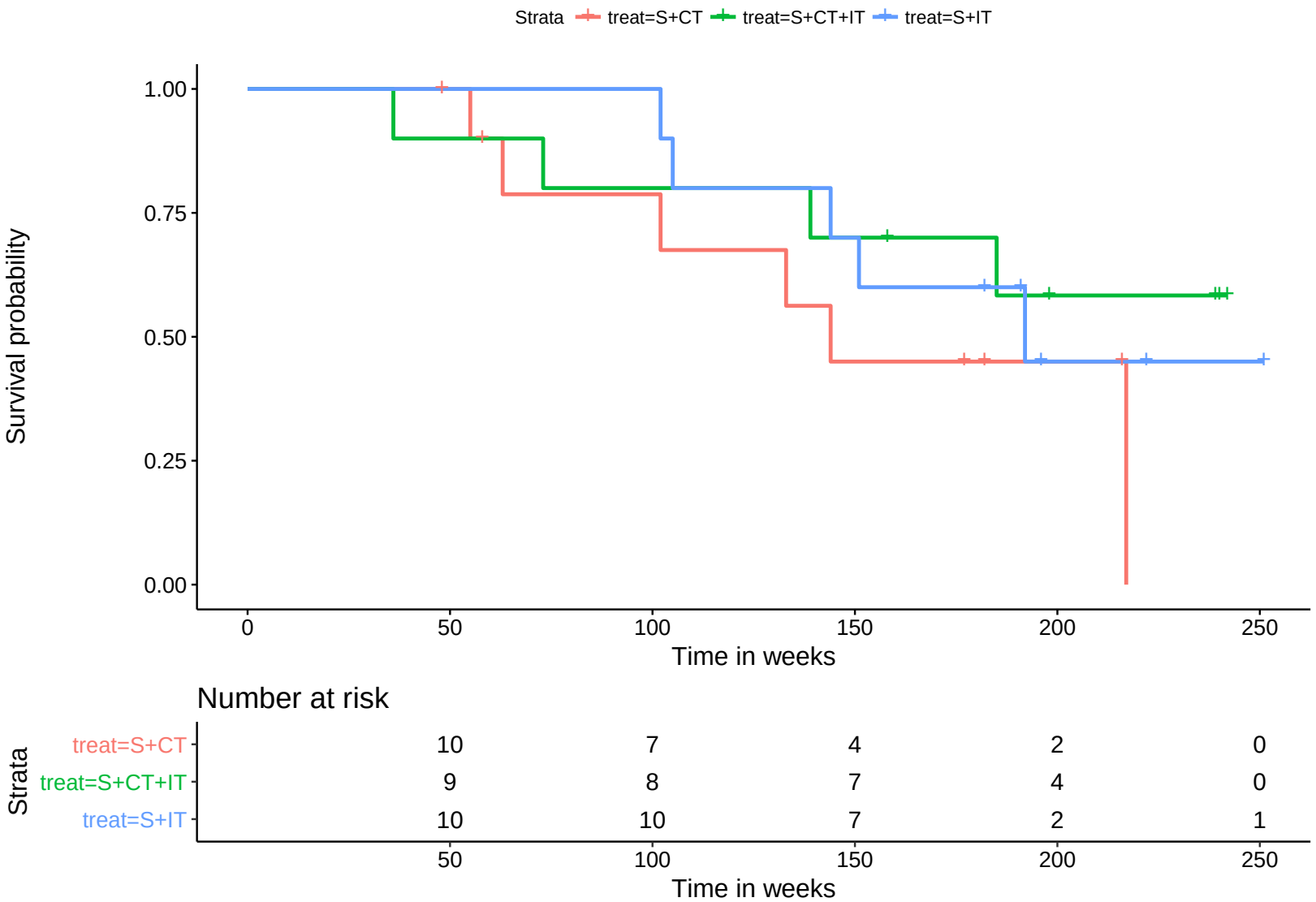
```
kmfit <- survfit(S ~ treat, dat = brca)

print(kmfit, print.rmean = TRUE)
```

```
Call: survfit(formula = S ~ treat, data = brca)

                n events *rmean *se(rmean) median 0.95LCL
treat=S+CT     11      6    153       21.1    144     102
treat=S+CT+IT  10      4    188       23.7     NA     139
treat=S+IT     10      5    188       17.9    192     144
               0.95UCL
treat=S+CT          NA
treat=S+CT+IT       NA
treat=S+IT          NA
    * restricted mean with upper limit =  242
```

summary(kmfit)

```
> summary(kmfit)
Call: survfit(formula = S ~ treat, data = brca)

                treat=S+CT
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   55     10       1    0.900  0.0949        0.732        1.000
   63      8       1    0.787  0.1340        0.564        1.000
  102      7       1    0.675  0.1551        0.430        1.000
  133      6       1    0.562  0.1651        0.316        1.000
  144      5       1    0.450  0.1660        0.218        0.927
  217      1       1    0.000     NaN           NA           NA

                treat=S+CT+IT
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   36     10       1    0.900  0.0949        0.732            1
   73      9       1    0.800  0.1265        0.587            1
  139      8       1    0.700  0.1449        0.467            1
  185      6       1    0.583  0.1610        0.340            1

                treat=S+IT
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  102     10       1     0.90  0.0949        0.732        1.000
  105      9       1     0.80  0.1265        0.587        1.000
  144      8       1     0.70  0.1449        0.467        1.000
  151      7       1     0.60  0.1549        0.362        0.995
  192      4       1     0.45  0.1743        0.211        0.961
```

K-M Plot via survminer



K-M Plot via survminer (code)

```
ggsurvplot(kmfit, data = brca,
           risk.table = TRUE,
           risk.table.height = 0.25,
           xlab = "Time in weeks")
```

Testing the difference between curves

```
survdiff(S ~ treat, dat = brca)
```

```
Call:
survdiff(formula = S ~ treat, data = brca)

                N Observed Expected (O-E)^2/E (O-E)^2/V
treat=S+CT     11        6     3.80    1.2772    1.7647
treat=S+CT+IT 10        4     5.62    0.4676    0.7725
treat=S+IT     10        5     5.58    0.0605    0.0981

 Chisq= 1.9  on 2 degrees of freedom, p= 0.393
```

What do we conclude?

Fit Cox Model A: Treatment alone

```
modA <- coxph(S ~ treat, data = brca)
modA
```

```
Call:
coxph(formula = S ~ treat, data = brca)

                coef exp(coef) se(coef)     z    p
treatS+CT+IT -0.831     0.435    0.655 -1.27 0.20
treatS+IT    -0.583     0.558    0.609 -0.96 0.34

Likelihood ratio test=1.75  on 2 df, p=0.416
n= 31, number of events= 15
```

summary(modA)

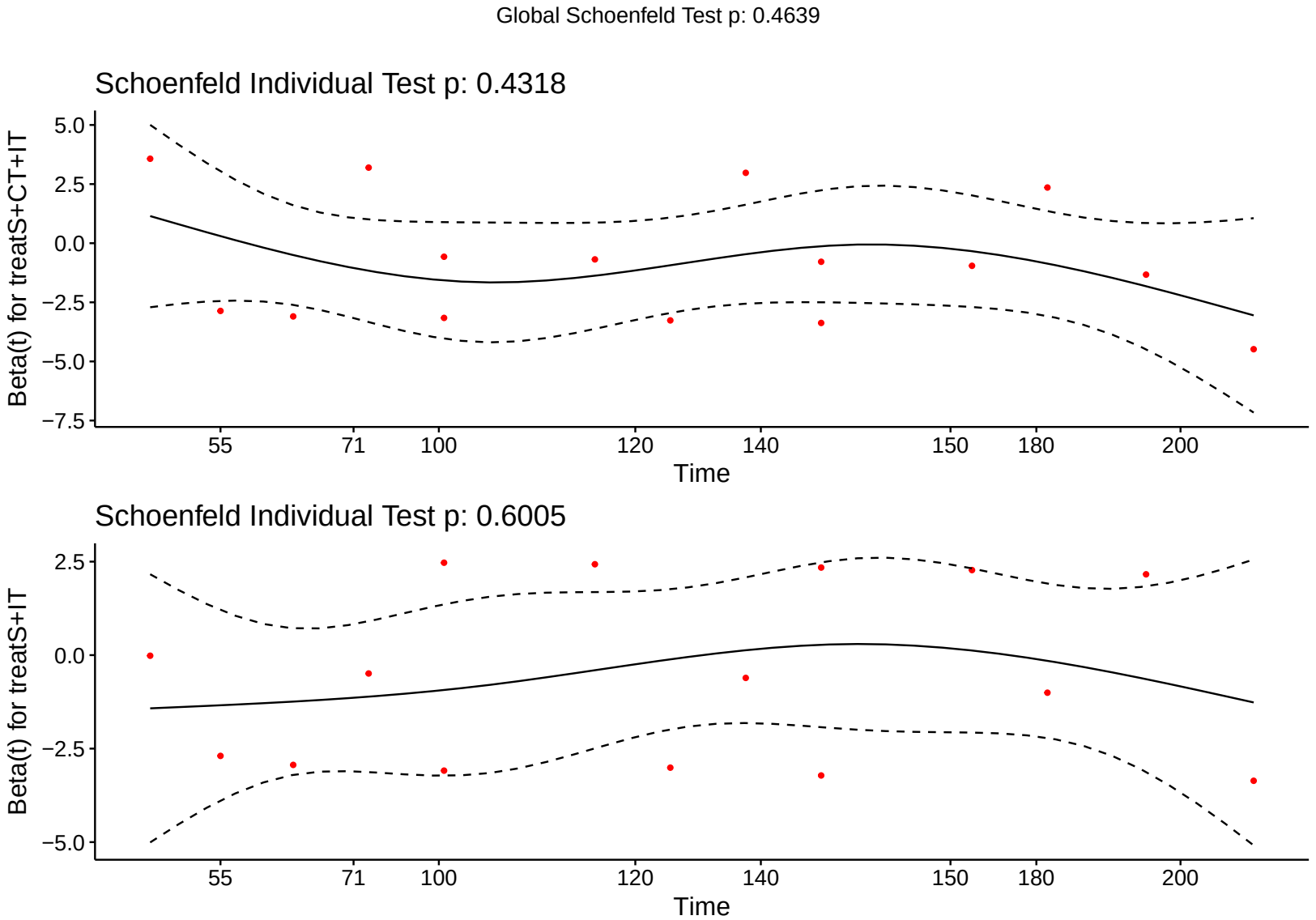
```
> summary(modA)
Call:
coxph(formula = S ~ treat, data = brca)

  n= 31, number of events= 15

                coef exp(coef) se(coef)      z Pr(>|z|)
treatS+CT+IT -0.8313    0.4355   0.6547 -1.270    0.204
treatS+IT    -0.5832    0.5581   0.6088 -0.958    0.338

             exp(coef) exp(-coef) lower .95 upper .95
treatS+CT+IT    0.4355      2.296    0.1207     1.571
treatS+IT       0.5581      1.792    0.1692     1.840

Concordance= 0.577  (se = 0.078 )
Rsquare= 0.055   (max possible= 0.944 )
Likelihood ratio test= 1.75  on 2 df,   p=0.4164
Wald test            = 1.82  on 2 df,   p=0.403
Score (logrank) test = 1.89  on 2 df,   p=0.3878
```

Check Proportional Hazards Assumption

```
cox.zph(modA)
```

```
                rho chisq     p
treatS+CT+IT -0.198 0.618 0.432
treatS+IT     0.138 0.274 0.601
GLOBAL           NA 1.536 0.464
```

Graphical PH Test `ggcoxzph(cox.zph(modA))`



Fit Cox Model B: Treatment + Age

```
modB <- coxph(S ~ treat + age, data = brca)
modB
```

```
Call:
coxph(formula = S ~ treat + age, data = brca)

                coef exp(coef) se(coef)     z     p
treatS+CT+IT -0.5996    0.5490   0.6574 -0.91 0.362
treatS+IT    -0.3116    0.7323   0.6094 -0.51 0.609
age           0.0781    1.0812   0.0367  2.13 0.034

Likelihood ratio test=6.99  on 3 df, p=0.0722
n= 31, number of events= 15
```

summary(modB)

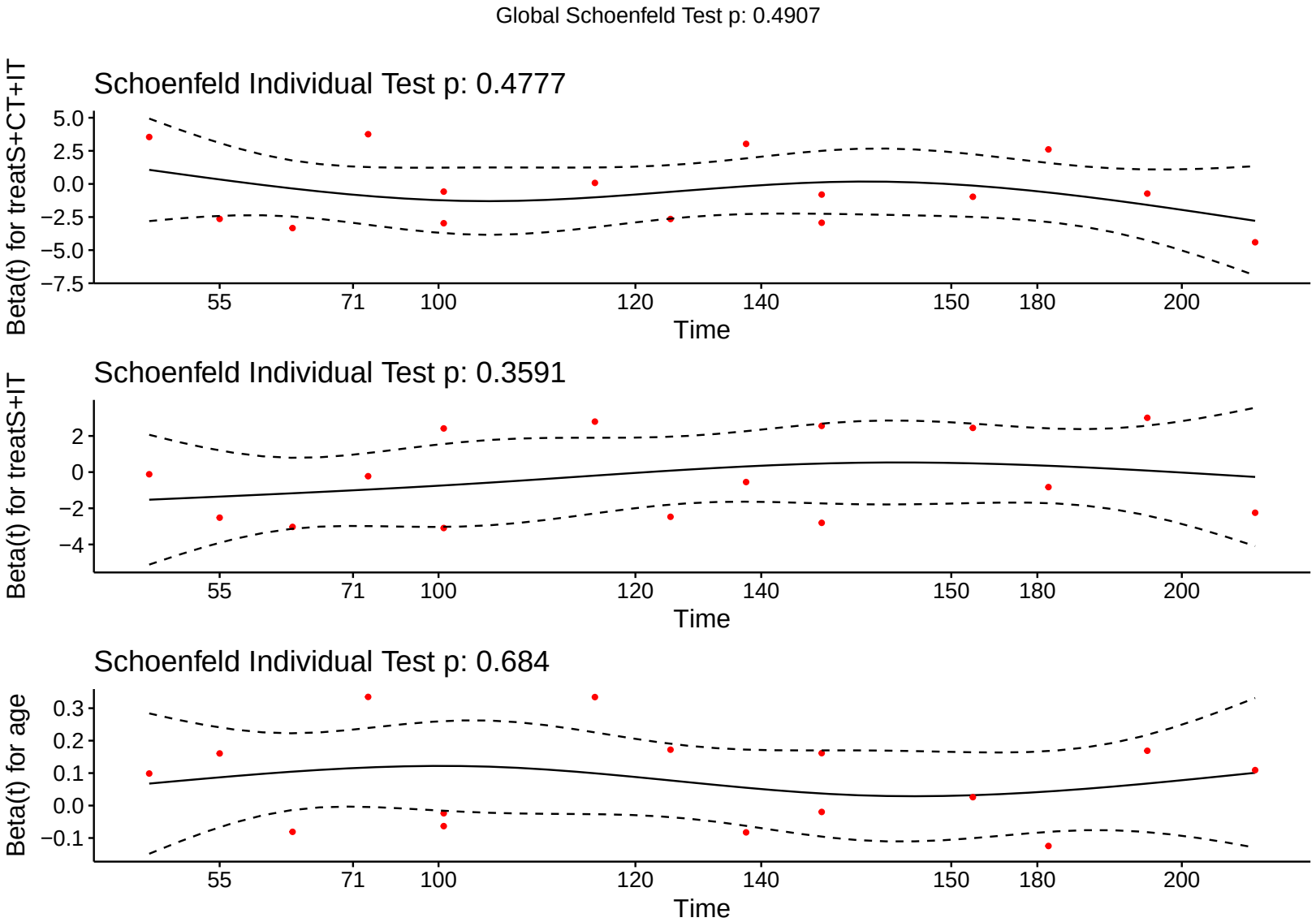
```
> summary(modB)
Call:
coxph(formula = S ~ treat + age, data = brca)

  n= 31, number of events= 15

                coef exp(coef) se(coef)      z Pr(>|z|)
treatS+CT+IT -0.59960   0.54903  0.65741 -0.912   0.3617
treatS+IT    -0.31161   0.73227  0.60936 -0.511   0.6091
age           0.07807   1.08119  0.03672  2.126   0.0335 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

             exp(coef) exp(-coef) lower .95 upper .95
treatS+CT+IT    0.5490     1.8214    0.1514     1.992
treatS+IT       0.7323     1.3656    0.2218     2.417
age             1.0812     0.9249    1.0061     1.162

Concordance= 0.701  (se = 0.083 )
Rsquare= 0.202   (max possible= 0.944 )
Likelihood ratio test= 6.99  on 3 df,   p=0.07224
Wald test            = 5.85  on 3 df,   p=0.1192
Score (logrank) test = 6.15  on 3 df,   p=0.1043
```

Proportional Hazards Assumption: Model B Check

```
cox.zph(modB)
```

```
                rho chisq     p
treatS+CT+IT -0.179 0.504 0.478
treatS+IT     0.244 0.841 0.359
age          -0.106 0.166 0.684
GLOBAL           NA 2.416 0.491
```

Graphical PH Test `ggcoxzph(cox.zph(modB))`



What to do if the PH assumption is violated

- If the PH assumption fails on a categorical predictor, fit a Cox model stratified by that predictor (use `strata(var)` rather than `var` in the specification of the `coxph` model.)
- If the PH assumption is violated, this means the hazard isn't constant over time, so we could fit separate Cox models for a series of time intervals.
- Use an extension of the Cox model that permits covariates to vary over time.

Visit https://cran.r-project.org/web/packages/survival/vignettes/timedep.pdf for details on building the relevant data sets and models, with examples.

Some Loose Ends

On Fitting Regression Models

Some things are always true. . .

- Being honest about your findings is important.
- Identifying a stable phenomenon to study is crucial.
- Measuring that phenomenon well is crucial.
- Being transparent about your work is important. Describing your work so it can be replicated, and then actually replicating it, are good things. Sharing your data and your materials is a good idea.
- Having a large sample size (n) is helpful in fitting models.

but the impact of lots of other things changes depending on why you're fitting a regression model.

The Key Point

Decide what your research question is, and use it to help you think about what's important in your modeling.

- Models that account for only a few of the possibly important dimensions of a problem don't lead to causal conclusions, but can help screen out important from less important areas for future work.
- Model development strategies that work well with large sample sizes (n) and small numbers of predictors to consider don't necessarily work well when the situation is reversed.
- Most problems involve missing data, and problems in measurement. Getting those issues settled effectively is often overlooked.
- The sample size you need to fit a regression model changes depending on your aims.

On Fitting Models

What are some of the reasons you might fit a linear (or generalized linear) model?

1. (All prediction, no explanations) Because you want to make predictions about an outcome in new data based on some training data you have, but you're happy to take those predictions as emerging from a mysterious magic "black box" that cannot be peered into without spoiling the surprise. You don't care if your results are a little biased, so long as the predictions are strong. Parsimony doesn't matter to you.

- Some especially useful tools here include: variable (feature) selection through cross-validation, stepwise approaches, AIC and BIC, machine learning tools like regression trees, and other means of quickly searching through many possible models.
- Sample size is rarely a big issue here. The big problem is having more variables than you can possibly plot at once. You usually have enough data to partition into separate development and test samples.

On Fitting Models

2. (All description, external validity is irrelevant) Because you want to describe, as accurately as possible, the nature of the associations you observe in the available data, but you don't care much at all whether the conclusions you draw will hold up in new data.

- Confidence intervals for coefficients (slopes, mostly), and sometimes you'll run the model on clinically relevant cutpoints rather than continuous predictors to see what's happening more simply. Simple polynomial models can be appealing, and you'll sometimes want to build this in the ANCOVA context, where you're looking for the impact of specific pre-specified interactions.
- Residual plots play a big role here in deciding whether the model "fits" well enough, or identifying cases when it doesn't.
- Cross-validation is useful, but not a big part of model selection or convincing people that the model is "right" or not.

On Fitting Models

3. (Clinical prediction) You want to do an excellent job predicting an outcome in new data, but you have a lot of prior knowledge about the predictors under consideration, and want to use that information to help produce prediction rules as effectively as possible. You welcome the fact that most relationships are non-linear, but would like to be parsimonious if possible, as data are often expensive.

- Some especially useful tools here include: scatterplot matrices (when the number of predictors is modest), cross-validation, assessments of discrimination and calibration, Spearman's ρ^2 plot to point the way to non-linear terms that might be impactful if present, restricted cubic splines, polynomial functions, and graphical tools like nomograms
- Most stepwise tools aren't helpful here. We try to not "peek" at the outcome-predictor relationships to maintain unbiased estimates of the relationship without extensive validation.

On Fitting Models

4. (Above all else, simplicity) You want the problem to look like one in a statistics textbook, where everything is fit with the simplest possible model, where every term adds statistically significant predictive value, and where obtaining an unbiased estimate of the outcome is especially important. You still care a bit about what happens in new data, but you're mostly concerned about parsimonious model development.

- Some especially useful tools include best subsets, stepwise approaches, and methods for pruning a set of predictors with clustering or principal components analysis. These models usually make the (often incorrect) assumption that relationships are linear.
- Often this approach is used by people who are trying to pre-specify their entire model in advance, and want to be sure they can "explain" the result when they are done. That may not be a reasonable thing to hope for. This is a place where the "rule of 20" can be very helpful in setting expectations in advance.

On Fitting Models

5. (Causal inference) You want to identify whether a particular causal pathway you have pre-specified matches up well with what you see in new data.
6. (Risk Adjustment) You want to identify the impact of a particular exposure/predictor on an outcome, while controlling for the effects of a series of additional predictors. Perhaps you've done a randomized experiment / clinical trial, and want to identify whether particular results meet a standard for statistical (as well as clinical) significance. Power is very important.

- In either case, bias is very important, and you want to avoid it. Careful design of a comparison group (like I teach in 500) is a very good way to go about this work, but it's also true that there's a lot of epidemiology that goes into drawing causal conclusions, or even thinking hard about an association.
- Often the details of modeling take a back seat here to the details of designing the study (and the comparison groups) in the first place.

Using Restricted Cubic Splines

Explaining a Model with a Restricted Cubic Spline

Restricted cubic splines are an easy way to include an explanatory variable in a smooth and non-linear fashion in your model.

- The number of knots, k, are specified in advance, and this is the key issue to determining what the spline will do. We could use AIC to select k, or follow the general idea that for small n, k should be 3, for large n, k should be 5, and so often k = 4.
- The location of those knots is not important in most situations, so R places knots by default where the data exist, at fixed quantiles of the predictor's distribution.
- The "restricted" piece means that the tails of the spline (outside the outermost knots) behave in a linear fashion.

The "Formula" from a Model with a Restricted Cubic Spline

- The best way to demonstrate what a spline does is to draw a picture of it. When in doubt, do that: show us how the spline affects the predictions made by the model.
- But you can get a model equation for the spline out of R (heaven only knows what you would do with it.) Use the `latex` function in the `rms` package, for instance.

An Example

```
d <- datadist(iris)
options(datadist = "d")
m1 <- ols(Sepal.Length ~ rcs(Petal.Length, 4) + Petal.Width,
          data = iris, x = TRUE, y = TRUE)
m1
```

```
Linear Regression Model

 ols(formula = Sepal.Length ~ rcs(Petal.Length, 4) + Petal.Wid
     data = iris, x = TRUE, y = TRUE)

                 Model Likelihood     Discrimination
                    Ratio Test           Indexes
 Obs      150    LR chi2    253.23    R2        0.815
 sigma0.3609    d.f.            4    R2 adj    0.810
 d.f.     145    Pr(> chi2) 0.0000    g         0.844
```

Function(m1)

```
Function(m1)
```

```
function (Petal.Length = 4.35, Petal.Width = 1.3)
{
    4.7226352 + 0.24335435 * Petal.Length + 0.021780541 * pmax
        1.3, 0)^3 - 0.037888523 * pmax(Petal.Length - 3.33, 0)
        0.00031123969 * pmax(Petal.Length - 4.8, 0)^3 + 0.0157
        pmax(Petal.Length - 6.1, 0)^3 - 0.33400958 * Petal.Wid
}
<environment: 0x000000002a0582c0>
```

What's in `Function(m1)`?

```
4.72 + 0.243  * Petal.Length
     + 0.022  * pmax( Petal.Length-1.3,  0)^3
     - 0.038  * pmax( Petal.Length-3.33, 0)^3
     + 0.0003 * pmax( Petal.Length-4.8,  0)^3
     + 0.016  * pmax( Petal.Length-6.1,  0)^3
     - 0.334  * Petal.Width
```

where `pmax` is the maximum of the arguments inside its parentheses.

Can we use "best subsets" or "all subsets" in logistic regression?

Using the `bestglm` package to do "best subsets" with logistic models

There are several available tools. If you're interested, I'd start with this link.

- The `bestglm` package does an exhaustive (all subsets) search through a glm (slowly) using AIC, BIC or cross-validation, for example.
- This expects the data to be in a certain form, for instance, the outcome must be named *y* and you can have no extraneous variables present.
- The tutorial you'll find at the link above is pre-tidyverse. I don't know how well `bestglm` works with the tidyverse, but it;s not likely to be much worse than `leaps` does.
- In addition to `bestglm` there are several other appealing tools for looking at large numbers of potential models quickly.

Next Time

- Retrospective Power and why most smart folks avoid it
  - Type S and Type M error: Saying something more useful
- Replicable Research and the Crisis in Science
  - ASA Statement on P values
  - Is changing the p value cutoff the right strategy?
  - Second-generation p values: A next step?