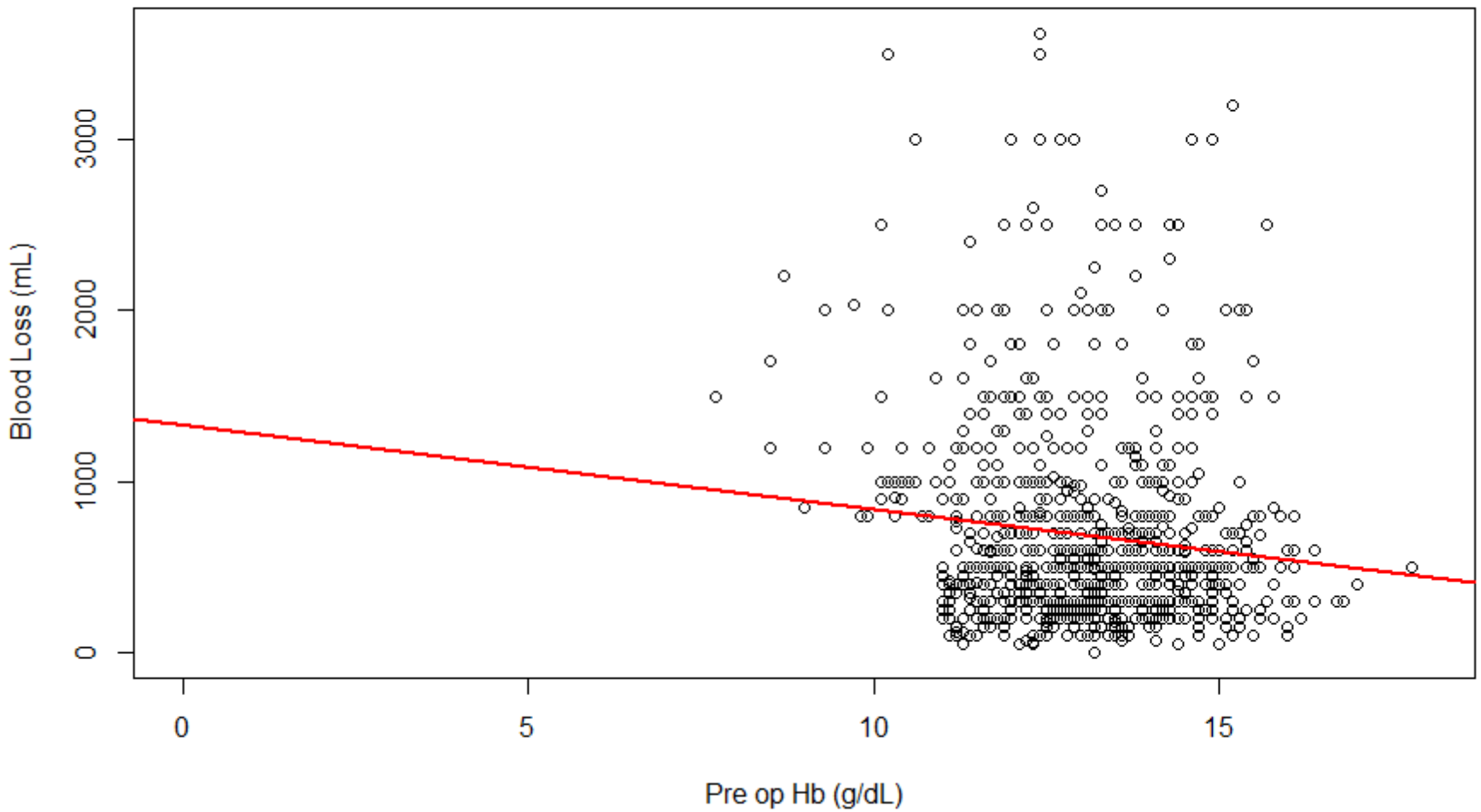
Biostatistics

Mithat Gonen

Brendon Bready

Intercept and Slope

- Pred(Y) = a + b*X
  - When X = 0 → Pred(Y) = a + b*0 = a
  - Intercept (a) is the point where the line crosses the vertical axis (Y value for X = 0)
- Slope
  - For X: Pred(Y) = a + b*X
  - For X+1: Pred(Y) = a + b*(X+1)
  - The difference in Pred(Y) when X goes up by one unit is: a + b*(X+1) – (a + b*X) = a + b*X + b – a – b*x = b
  - Slope is the change in Y when X changes one unit


Blood Loss (mL)
0
1000
2000
3000
0
5
10
15
Pre op Hb (g/dL)

Parameter	Estimate	95% CI (Lower Bound)	95% CI (Upper Bound)	p
Intercept	1330.35	966.20	1694.51	<0.0001
Hb	-49.41	-77.10	-21.72	0.0005
R2	0.012			

Multivariable (multivariate) regression

- We have more than one X
- And some of these X's can be continuous, some can be categorical
- I will first add a categorical variable to the mix
- It is very very very important to write the regression equation every time
- X1 is continuous, X2 is binary
- Pred(Y) = a + b1*X1 + b2*X2

Continuous and binary variables in regression

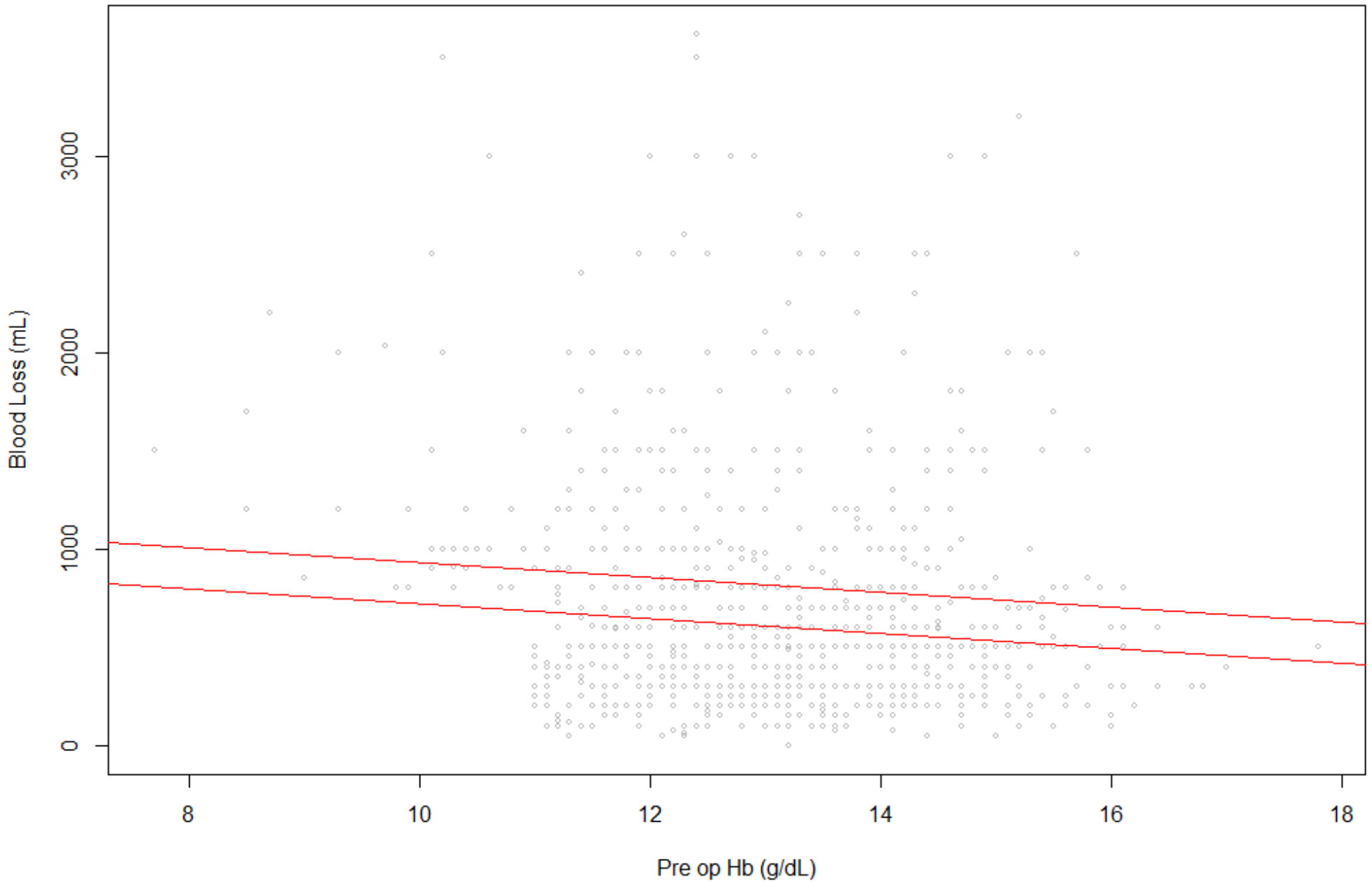
- Pred(Y) = a + b1*X1 + b2*X2
- Remember X2 is binary, so either 0 or 1
- When X2 = 0
  - Pred(Y) = a + b1*X1
- When X2 = 1
  - Pred(Y) = a + b1*X1 + b2
  - Pred(Y) = (a + b2) + b1*X1

Continuous and binary variables in regression

- Pred(Y) = a + b1*X1 + b2*X2
- Remember X2 is binary, so either 0 or 1
- When X2 = 0
  - Pred(Y) = a + b1*X1
- When X2 = 1
  - Pred(Y) = a + b1*X1 + b2
  - Pred(Y) = (a + b2) + b1*X1

Continuous and binary variables in regression

- Pred(Y) = a + b1*X1 + b2*X2
- X2 = 0 → Pred(Y) = a + b1*X1
- When X2 = 1 → Pred(Y) = (a + b2) + b1*X1
- Same slope, different intercept
- Parallel lines


Blood Loss (mL)
0
1000
2000
3000
8
10
12
14
16
18
Pre op Hb (g/dL)

Parameter	Estimate	95% CI (Lower Bound)	95% CI (Upper Bound)	p
Intercept	1103.69	734.50	1472.89	<0.0001
Hb	-37.93	-65.56	-10.29	0.0071
Binary Variable	206.26	127.55	284.98	<0.0001
R2	0.041			

Pred(Y) for X2 = 0 → 1103.69 – 37.93*Hb

Pred(Y) for X2 = 1 → 1309.95 – 37.93*Hb

What if we want different slopes

- We need to use an interaction
- Pred(Y) = a + b1*X1 + b2*X2 + b12*X1*X2
- When X2 = 0
  - Pred(Y) = a + b1*X1
- When X2 = 1
  - Pred(Y) = (a + b2) + (b1 + b12)*X1
- Different intercepts and slopes

Parameter	Estimate	95% CI (Lower Bound)	95% CI (Upper Bound)	p
Intercept	864.78	501.49	1428.06	<0.0001
Hb	-27.44	-62.23	7.35	0..122
Binary Variable	574.90	-171.88	1321.68	0.131
Hb*Binary Variable	-28.44	-85.73	28.86	0.330

Pred(Y) for X2 = 0 → 864.78 – 27.44*Hb

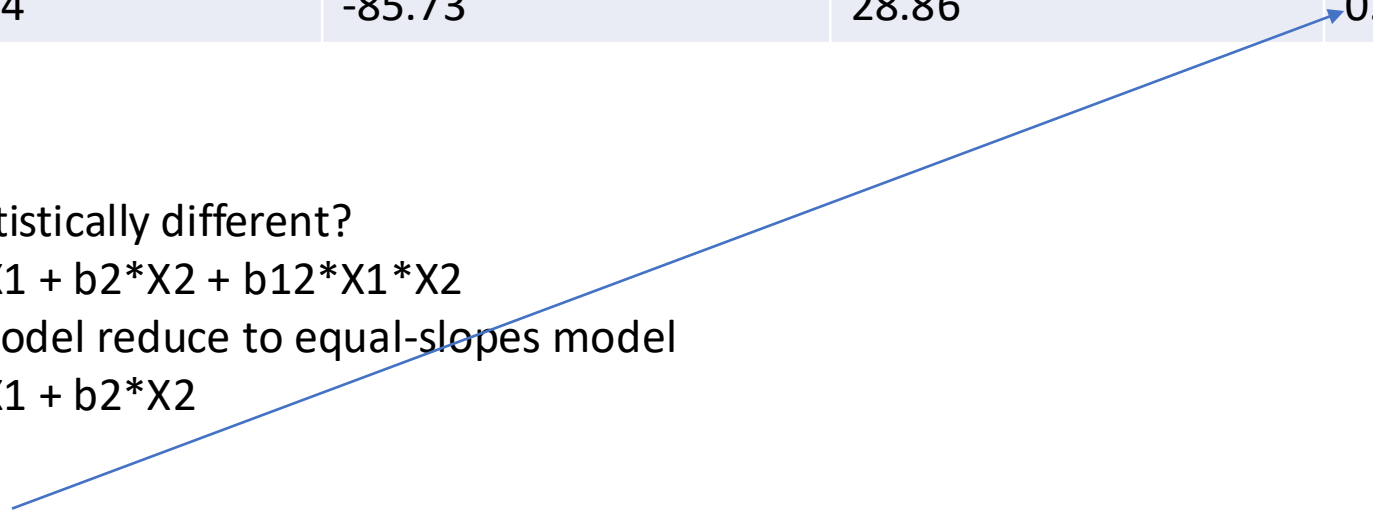
Pred(Y) for X2 = 1 → 1449.68 – 55.88*Hb

Parameter	Estimate	95% CI (Lower Bound)	95% CI (Upper Bound)	p
Intercept	864.78	501.49	1428.06	<0.0001
Hb	-27.44	-62.23	7.35	0..122
Binary Variable	574.90	-171.88	1321.68	0.131
Hb*Binary Variable	-28.44	-85.73	28.86	0.330

Are the slopes statistically different?
Pred(Y) = a + b1*X1 + b2*X2 + b12*X1*X2
When does this model reduce to equal-slopes model
Pred(Y) = a + b1*X1 + b2*X2

Parameter	Estimate	95% CI (Lower Bound)	95% CI (Upper Bound)	p
Intercept	864.78	501.49	1428.06	<0.0001
Hb	-27.44	-62.23	7.35	0..122
Binary Variable	574.90	-171.88	1321.68	0.131
Hb*Binary Variable	-28.44	-85.73	28.86	0.330

Are the slopes statistically different?
Pred(Y) = a + b1*X1 + b2*X2 + b12*X1*X2
When does this model reduce to equal-slopes model
Pred(Y) = a + b1*X1 + b2*X2
When b12 = 0!
So test for b12 = 0



Adding more variables

- Once you have two continuous predictors (X's) in the model it becomes difficult to visualize, more than 2 impossible
- So we need to "imagine" points in higher dimensional spaces that only exist in our minds (and in mathematical representations)
- But the idea of a "slope", "residual" etc applies

Parameter	Estimate	95% Lower	95% Upper	P
Intercept	1303.07	501.49	1428.06	<0.001
Hb	-37.52	-62.23	7.35	0.008
Binary Variable	207.473	-171.88	1321.68	<0.001
Age	-3.48	-85.73	28.85	0.018
R2	0.046			

Interpretation of coefficients

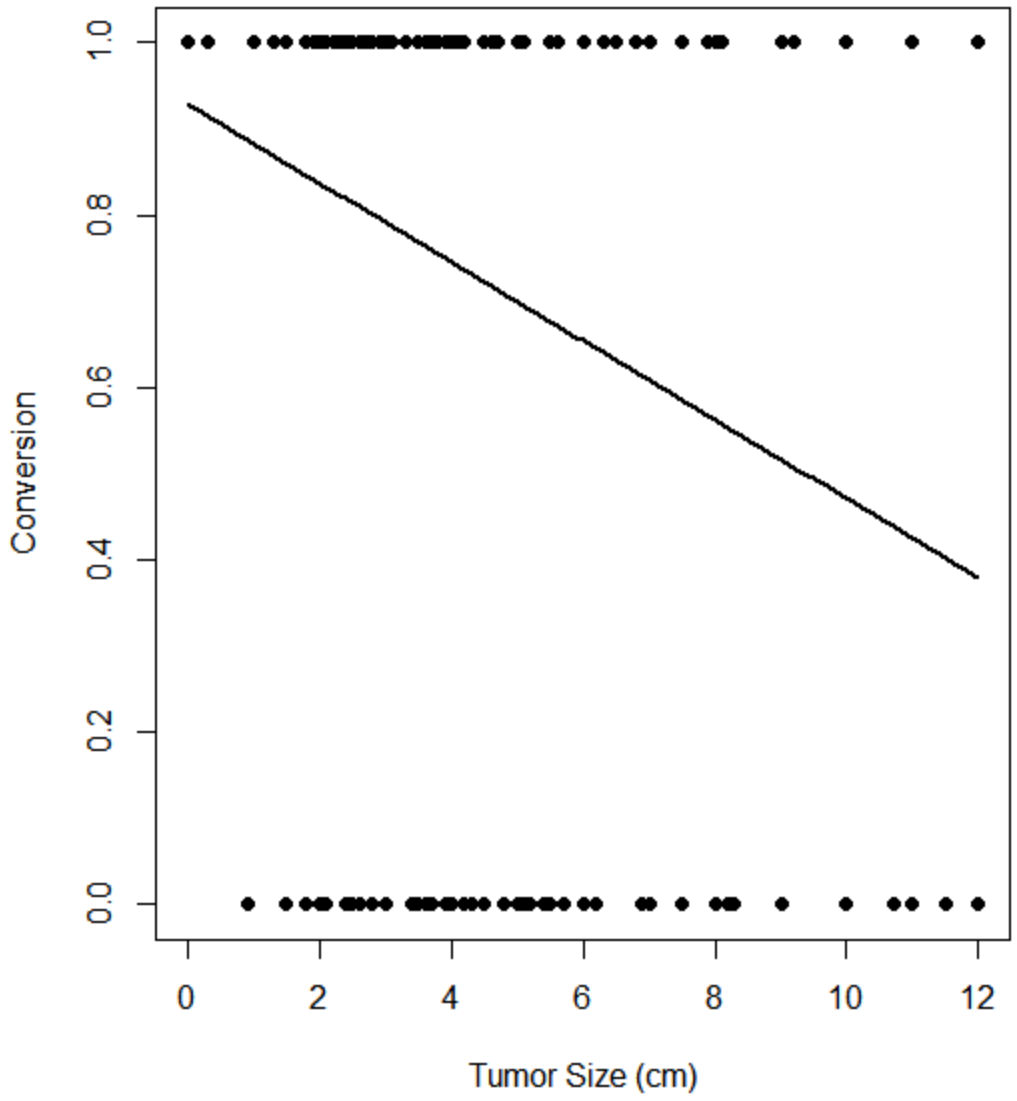
- Change in Pred(Y) corresponding to one unit change in X keeping others constant
- Binary: one unit means going from 0 to 1
- Continuous: depends on units of measurement
- Age: Take two patients with same Hb and X2, Pred(Y) goes down by 3.48 when age goes up by one year
- X2: Take two patients with same Hb and Age. The one with X2 = 1 has Pred(Y) higher by 207.47

Take a look at all the different models we fit

- Estimates jumped around quite a bit
- Model results are usually sensitive to which variables are included and excluded
- Variable selection is the Achilles heel of multivariate regression
- More on this in next lectures

Binary Outcome

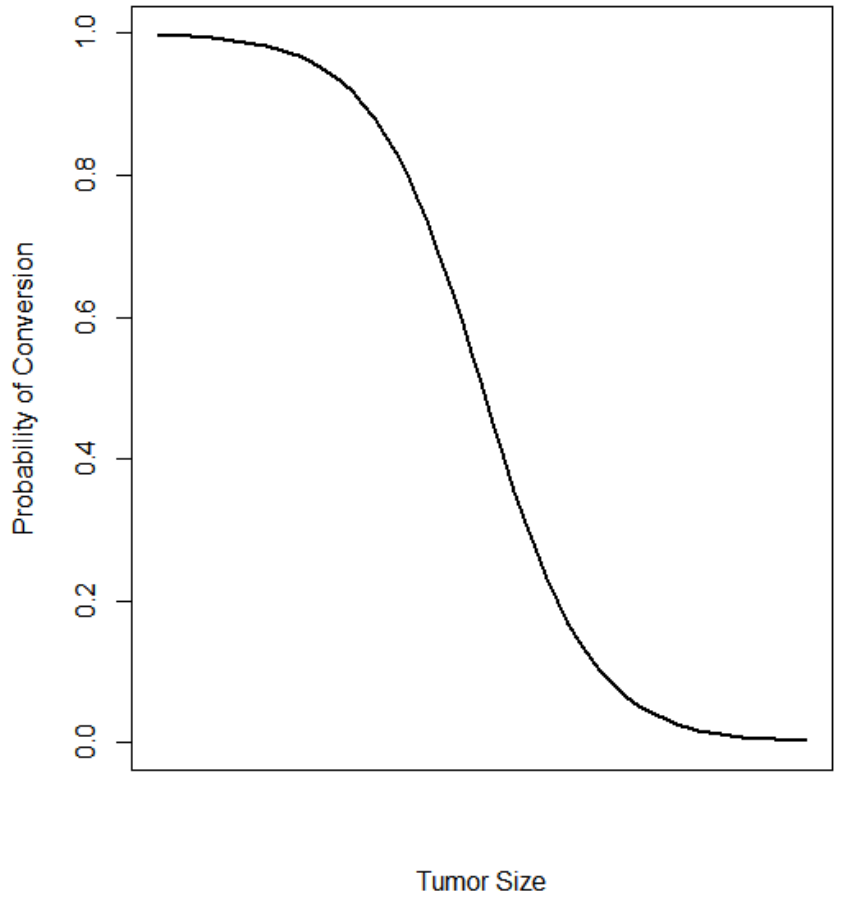
- Useful to code it as 0-1
- Example: NAC in breast cancer for conversion to being eligible for breast conservation surgery
- Start with one predictor: tumor size (cm)
- We coded the outcome as 0-1, can we use linear regression?
- Conversion = a + b*TumorSize



- What does a 0.8 for conversion mean?
- If you extend the line it will cross to negative at a large value of tumor size, what does it mean to have a negative value for conversion?

We Need to Model Probabilities

- Prob (Conversion) = a + b*TumorSize ??
- Might imply probabilities > 1 or < 0
- What we need is a sigmoid
  - Between 0 and 1 by definition



How to get a sigmoid?

- There are many many ways
- Let p = Prob(conversion)
- p/(1-p) = Odds Ratio
  - Odds are 3:1 means p =0.75
- log(p/(1-p)) = a + b*TumorSize → Logistic Regression
- p/(1-p) = exp(a + b*TumorSize) → another way to represent logistic regression
- p = exp(a + b*TumorSize)/(1 + exp(a + b*TumorSize)) → another way to represent logistic regression

Three Forms of Logistic Regression

- $\log(p/(1-p)) = a + b*TumorSize$ → Log Odds Ratio Version
- $p/(1-p) = \exp(a + b*TumorSize)$ → Odds Ratio version
- $p = \exp(a + b*TumorSize)/(1 + \exp(a + b*TumorSize))$ → Sigmoid version
- We use the log odds ratio version to estimate the parameters
- We use the odds ratio version to report (mostly, although sometimes the log odds ratio version is also used)
- We use the sigmoid version to plot

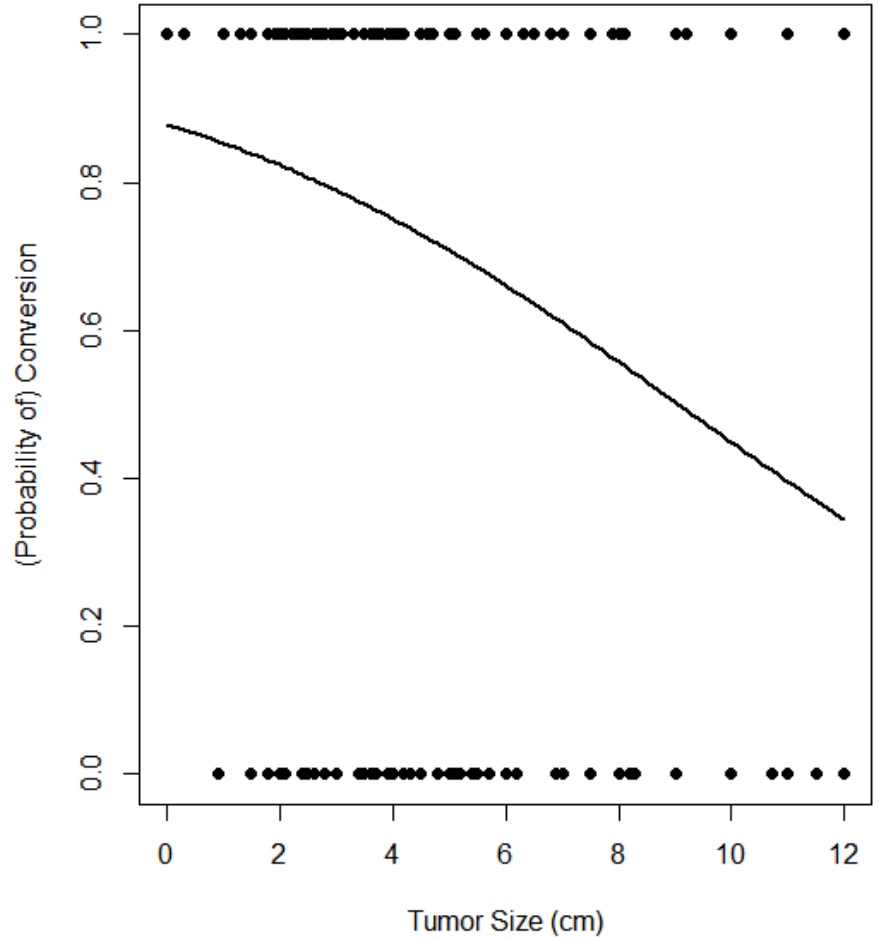
Three Forms of Logistic Regression

- We use the log odds ratio version to estimate the parameters
- We use the odds ratio version to report (mostly, although sometimes the log odds ratio version is also used)
  - You should be able to look at a paper and tell which version they are reporting
- We use the sigmoid version to plot

Logistic Regression: Estimate

- log(p/(1-p)) = a + b*TumorSize
- Left hand side is the “log odds ratio”
- Right hand side is the “linear predictor”
- We do not have a linear regression but we have a linear predictor
- Or we have a regression linear in log odds ratio.

Logistic Regression: Plotting



Logistic Regression: Interpretation

- log(p/(1-p)) = 1.97 - 0.22 * Tumor Size
- How do we interpret the coefficients?
- 1.97 is the intercept of the log-odds
  - When Tumor Size is 0, log odds is 1.97
- -0.22 is the slope of the log-odds
  - For each cm increase in tumor size log odds ratio decreases by 0.22
- It is not so practical with the log in there

Logistic Regression: Interpretation

- (p/(1-p)) = exp(1.97 - 0.22 * Tumor Size)
- Let's interpret the "slope" first
- What happens when you increase tumor size by 1 cm
  - exp(1.97 - 0.22 * (Tumor Size + 1))
  - exp(1.97 - 0.22 * Tumor Size)
  - The ratio of the right-hand sides is exp(-0.22) = 0.80
- One cm increase in Tumor size multiplies the odds by 0.80
- Or one cm increase in Tumor size decreases the odds by 20%

Logistic Regression: Intercept

- Not commonly needed to interpret but I will tell you anyway
- (p/(1-p)) = exp(1.97 - 0.22 * Tumor Size)
- When Tumor Size is 0 odds (of response) is exp(1.97) = 7.22
- p/(1-p) = 7.22 → p = 0.88

Variable	Odds Ratio	95% Lower Bound	95% Upper Bound	p
Intercept	7.22	4.57	11.25	<0.001
Tumor Size (cm)	0.80	0.73	0.88	<0.001

Tumor size is significantly associated with probability of conversion (Odds Ratio: 0.80, 95%CI: 0.73 – 0.88, p<0.001)

One cm increase in tumor size reduces the odds of conversion by 20%. We are 95% confident that the odds of conversion are reduced by a factor of 12% to 27%. Where did these numbers come from? 1-0.8, 1-0.73, 1-0.88.

It is also common to say one cm increase in tumor size reduces the probability of conversion by 20%. Incorrect but commonly used.

Variable	Odds Ratio	95% Lower Bound	95% Upper Bound	p
Intercept	7.22	4.57	11.25	<0.001
Tumor Size (cm)	0.80	0.73	0.88	<0.001

Tumor size is significantly associated with probability of conversion (Odds Ratio: 0.80, 95%CI: 0.73 – 0.88, $p<0.001$)

What is the null hypothesis? $b = 0$ or $\exp(b) = 1 \rightarrow OR = 1$

Binary Covariate

- Example: Clinical N+
- Equation: log(p/(1-p)) = a + b*NodePos
- Estimates
  - a = 0.81
  - b = -0.15
- Interpret the “b”

Remember to use the odds ratio version for interpretation

- $p/(1-p) = \exp(0.81 - 0.15 * \text{NodePos})$
- NodePos can only take two values: 0 (negative), 1 (positive)
- Therefore a one unit increase represents the odds ratio for positives to negatives
- $\exp(-0.15) = 0.86$ is the odds ratio for NodePos
- Node positivity decreases the odds of conversion by 14%

Building a Logistic Regression Model

- Let’s use both tumor size and node positivity in the same model
- Log(p/(1-p)) = a + b*TumorSize + c*NodePos

Variable	Odds Ratio	95% Lower Bound	95% Upper Bound	p
Tumor Size (cm)	0.96	0.94	0.98	<0.001
Node +	0.87	0.81	0.94	<0.001

It is very very very important to write the regression equation every time

$$\text{Log}(p/(1-p)) = a + b*\text{TumSize} + c*\text{NodePos}$$

Interpretations

- When tumor size increases by 1cm and nodal status is held constant the odds of conversion decreases by 4%
- Helpful to think of two patients: same nodal status but tumor sizes differ by 1cm
- The one with the larger tumor size is 4% less likely to convert
- Same nodal status means they could both be negative, or they could both be positive, 4% applies regardless
  - “Equal slopes” assumption

Interpretations

- When tumor size is held constant nodal positivity reduces odds of conversion by 13%
- Again, think of two patients: same tumor size but one is N+ the other is N-
- The odds of conversion is 13% less for the N+ patient
- Again, applies to any tumor size because of “equal slopes”

Equal Slopes

- It is an assumption in this model … 4% applies to both N0 and N+ patients
- Log(p/(1-p)) = a + b*TumorSize + c*NodePos + d*TumorSize*NodePos
- A term that has the product of two variables is called an interaction
- Adding an interaction allows for different slopes
- How?

Different Slopes

- Log(p/(1-p)) = a + b*TumorSize + c*NodePos + d*TumorSize*NodePos
- What is the equation for N0 patients?
  - NodePos = 0 → Log(p/(1-p)) = a + b*TumorSize
- What is the equation for N1 patients
  - NodePos = 1 → Log(p/(1-p)) = (a + c) + (b + d)*TumorSize
- Slopes now differ by d

Effect of Tumor Size

- Odds ratio was 0.80 when we only used tumor size in the model
- It is 0.96 when it is used in conjunction with node positivity
- Which one is true?
- Achilles' heel of regression

All interpretations require a correctly-specified model

- Memorize this
- Interpretations depend on model being correctly specified
- Correctly specified model means all the important variables are included, all the unimportant ones excluded
- How do we ever know?

Correct Model

- Choosing the correct model is the most important thing in any regression application
- As much an art as it is science
- It is very important to pre-specify a list of candidate variables
- Then try a combination of these variables to find the best model empirically

Pre-specification

- Requires subject matter knowledge
- Beware the statistician or the computer scientist who says “just give me the data”

Finding the best model empirically

- An important area of statistics
- There is no one widely-accepted method
- Most common in medical research
  - Forward selection
  - Backward selection

Forward Selection

- Take your candidate variables
- For each one of them run a logistic regression with only that variable (called univariate analysis sometimes)
- Choose the variable with the smallest p
- Enter that into the model. Now run a logistic regression for all the remaining variables one by one. In each logistic regression include that variable and the variable already chosen to be in the model in the previous step
- Continue until you do not have $p<0.05$

Backward Selection

- Run a logistic regression with all your variables in the model
- IF all $p < 0.05$, then you have your model
- If not, exclude the variable with the highest p and re-run
- Continue until all $p<0.05$

Which one to use

- Backward selection requires more data points
- Rule of thumb: 10-15 “events” per variable in the regression model
- Event is the least common of the outcome categories
- 450 of 626 (72%) converted
- Non-conversion is the least common: so the number of events is 176
- One can include 12-17 variables here.
- Forward selection is less demanding but you have to stop adding to the model when you reach the rule of thumb.

What we mean by variable?

- When we say 10-15 events per variable
- We actually mean number of coefficients in the model, excluding the intercept (a)
- So an unequal slopes model has 3 “variables”

Back to Confounding

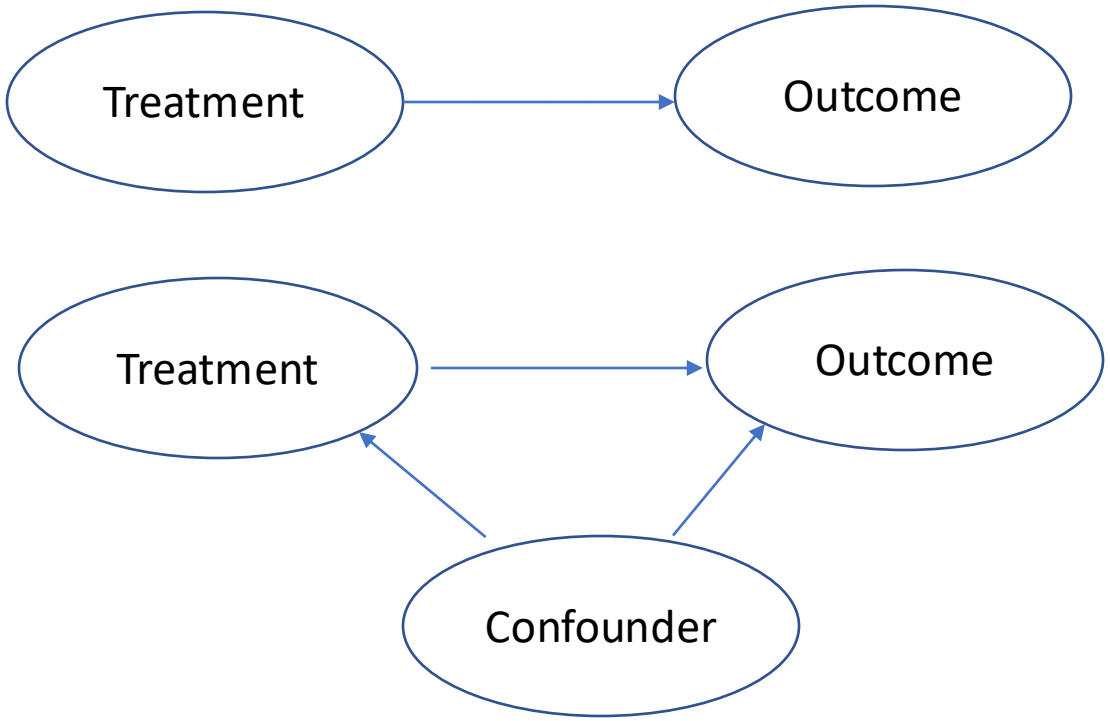
- We reviewed matching
- We saw how regression can be used
  - Put all your confounders and your treatment variable in a regression model
  - The interpretation of treatment variable is “hold all confounders constant, the treatment coefficient is the treatment effect estimate” adjusted for confoudners
- Propensity Score Matching

CY's Data

- Treatment: Embolization (n=64) vs Y90 (n=17)
- Confounders: Age, gender, histology, number of tumors, largest tumor size
- Hard to match on all of them even with large calipers or groupings

What Are We Trying to do?

- Remove the arrow from confounder to treatment



What if

- I am able to calculate the probability of receiving treatment (propensity) based on confounders
- Can I use it to match?
- Would it remove the link?
- How can I actually do it?

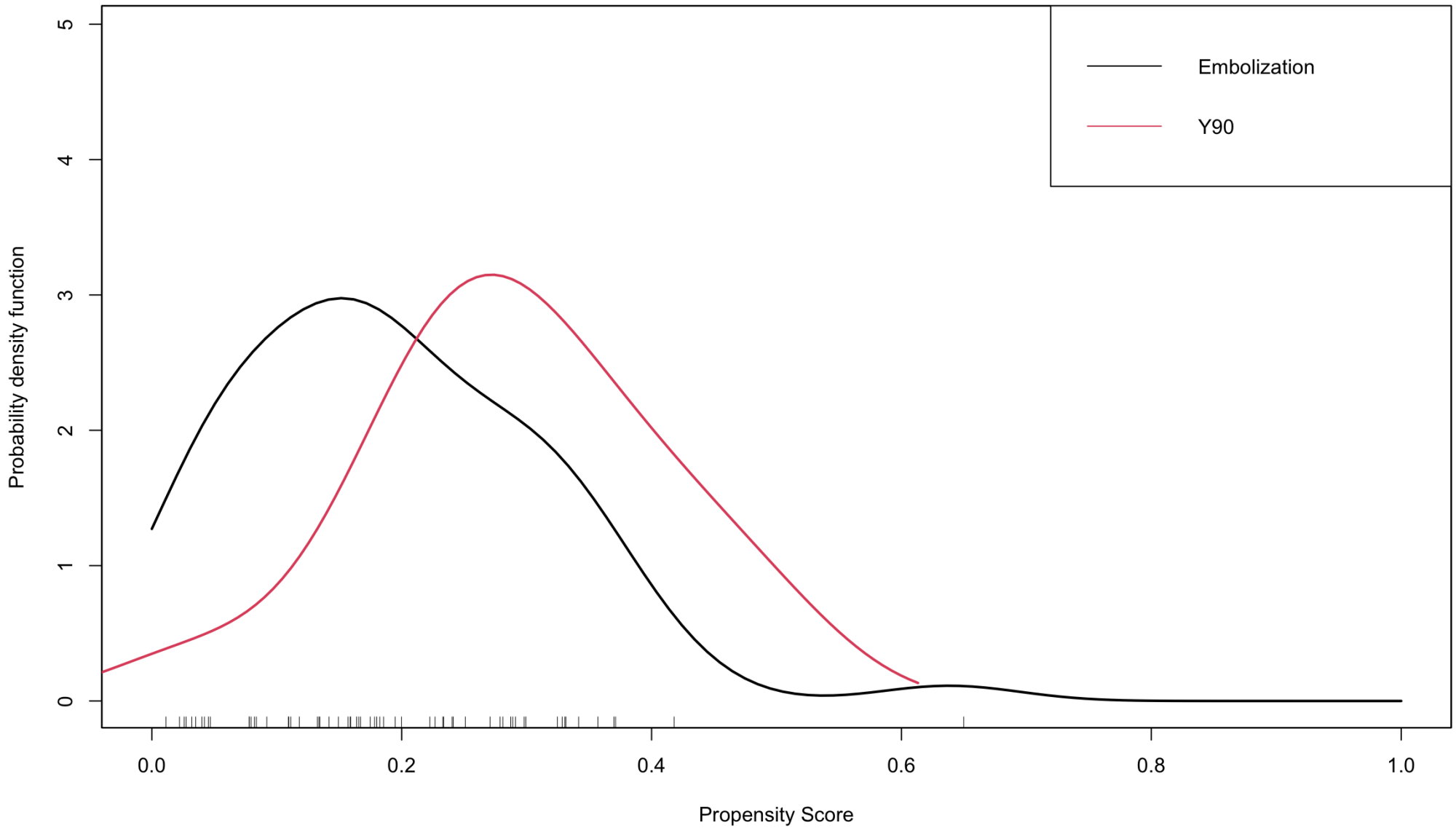
Logistic Regression?

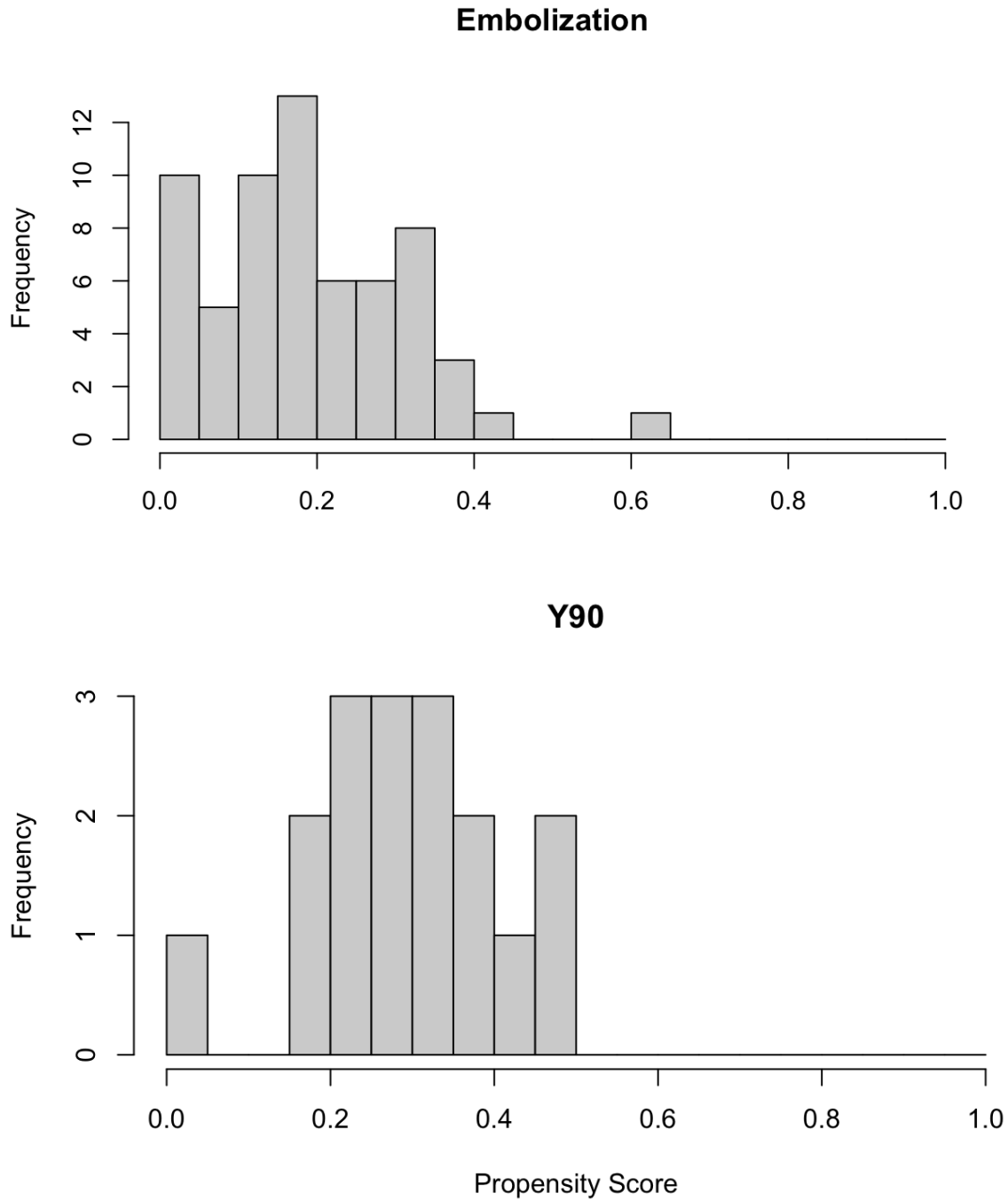
- Outcome (treatment): binary
- Fit a model and estimate the probability for each patient
- Match patients with similar probabilities
  - Side question: probabilities are continuous – how can we match?

Model

log(p/(1-p)) = -5.41 + 0.06*Age(years) – 0.016*Female +
0.18*Cutaneous – 0.29*Uveal - 0.39(2-5 Tumors) + 0.65*(5-10 Tumors)
– 0.06*(>10 Tumors) – 0.0034*Size(cm)

Let’s check the signs of coefficients. Do they make sense?


Embolization
Y90
Probability density function
0
1
2
3
4
5
0.0
0.2
0.4
0.6
0.8
1.0
Propensity Score


Embolization
Frequency
0
2
4
6
8
10
12
0.0
0.2
0.4
0.6
0.8
1.0
Y90
Frequency
0
1
2
3
0.0
0.2
0.4
0.6
0.8
1.0
Propensity Score

Overlap score

- C-index, AUC etc
- 0.5 (full overlap), 1 (none)
- We do not want either extreme (why?)
- In this case c = 0.73, sweet spot