1 Is there a difference between $\hat{y}$ and $\bar{y}$?

I refer to $\hat{y}_i$ as the predicted value from the fitted regression model for the ith observation. $\bar{y}$ is the sample mean.

2 What is the parameter we are trying to find in the ratio linear regression?

The ratio of means of the "exposed" group vs. the unexposed group.

3 Why are there t-statistics and F-statistics given as outputs of a regression model? Why are we analyzing t if the F is what is usually reported for regression models?

The t statistic is for each individual variable where $H_0 : \beta_j = 0$. The F statistic is for the overall fit of the model where $H_0 : \beta_1 = \beta_2 = \cdots = \beta_j = 0$ and the alternative is at least one β_j is not equal to 0. When there is only one predictor (determinant) variable, then $t^2_{statistic}$ is equal to $F_{statistic}$ as can be seen in example 2 in the 2nd regression handout.

4 Kind of confused about the overall goal of a regression... Is it just a faster way to get all the individual outputs (like the estimates, p-value, standard error, etc.), or does it have some extra use in and of itself?

It is useful to get proper standard errors, particularly when we want inference on the risk ratio. The benefits of regression will become even more evident when there is more than just one determinant. Regression summarizes the data into relatively few numbers represented by the regression estimates.

5 Why is south estimate log(theta hat), why isnt it log(u) + log (theta)

The model we are fitting is given by $\log(\mu) = \log(\mu_0) + \log(\theta) \times South$. Therefore the coefficient estimate for South is given by $\log(\theta)$.

6 Why is the df $n - p$ and not $n1 - 1 + n_2 - 1$?

The degrees of freedom for the standard deviation is $n1 - 1 + n_2 - 1$. The $n - p$ is for the model. We can think of it as a measure of how flexible our model is. If the degrees of freedom is small, then we have a very complicated model. Recall that p is the number of variables.

7 When calculating residual $(y_i - \hat{y}_i)$, is $\hat{y}_i$ the sample mean for the entire sample, or the sample mean for North and for South? And if it's the latter, are you subtracting $\hat{y}_i$ for North from all the North y's and $\hat{y}_i$ for South for all the South y's, or subtracting both $\hat{y}_i$ for each y?

- $\hat{y}_i$: is the predicted depth of the ocean based on the determinants of the model. This results in a vector of length n (i.e., it would include both north and south predictions in one long vector)
- y_i: is the observed depth of the ocean.
- the residual is simply the difference between these two vectors.