DUE: Wednesday, October 5, 11:59PM (34 points scaled to 100 percent)

You may discuss this assignment with whomever you wish, but please prepare and submit work in groups of ***ONE to THREE*** students, no more and no fewer. ***Each group*** will submit a copy of their group's completed assignment, via BbLearn, including the ***names and student ID numbers of all group members*** who participated on the assignment. If you discover a mistake, you may submit another version before the deadline. The last submitted (on-time) version will be graded, with all team members receiving the same score, which will be recorded in BbLearn.

GROUP MEMBERS WHO DO NOT CONTRIBUTE SUBSTANTIALLY TO AN ASSIGNMENT MAY BE REQUIRED TO WORK IN THEIR OWN GROUP OF ONE FOR THE REMAINDER OF THE SEMESTER.

While you are permitted to discuss the assignment with other groups, please prepare your own group's code/output and written answers. GROUPS WHOSE CODE AND SOLUTIONS APPEAR SUBSTANTIALLY SIMILAR MAY BE SUBJECT TO A 10% PENALTY.

Please prepare solutions in a ***neat, organized and concise fashion***! I prefer typeset presentations (e.g., cut and paste code/output into MS Word with added exposition when appropriate; knitr via EMACS and ESS; knitr or R Markdown via RStudio, the latter being the method preferred by students in recent years). At the very least, you need to ensure code and output are presented with a fixed-width font. Neatly handwritten presentations may also be appropriate for some problems. Sloppily prepared or disorganized solutions will not receive full credit.

To complete the items below, I expect you to find and use material in our lecture notes, including code/output, possibly after some modification. Some questions may be answered with code and output alone, but some exposition may be required beyond code and output for other questions. It's up to you to communicate concisely!

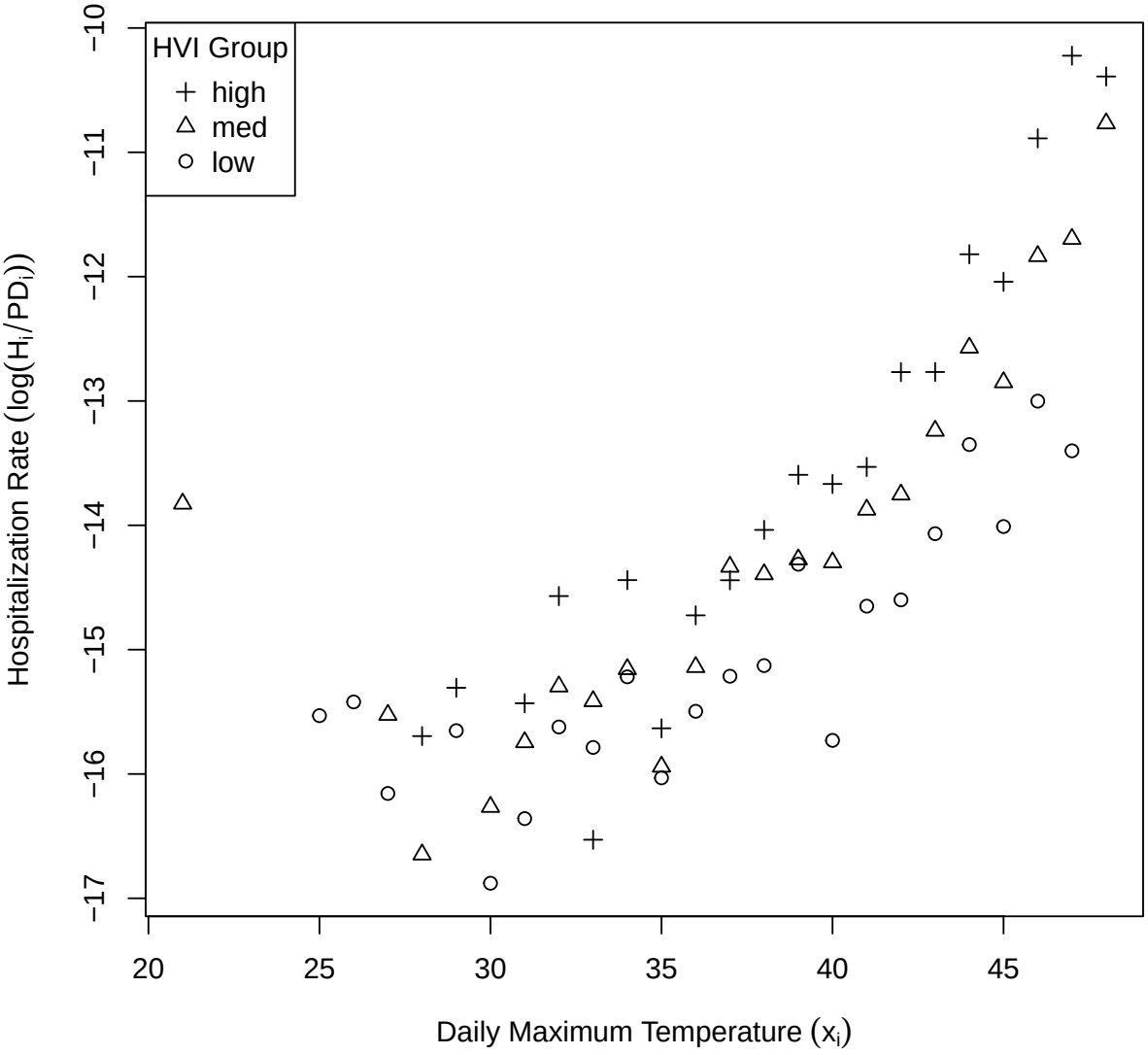
The data for this homework are part of a study, by Ben Ruddell, myself and others, on how the rate of heat-related hospitalization is related to temperature in a major metropolitan area and how this rate varies among groups classified according to their heat vulnerability.

- H_i, hospitalization count in area i (response)
- x_i, daily maximum temperature in area i (Celsius) (covariate)
- HVI_i, heat vulnerability index for area i (factor with three levels). We expect heat-related hospitalizations to increase as HVI goes from low to medium to high.
- PD_i, total person-days of exposure at temperature x_i in area i.
- As our response, we will work with (the natural log of) the rate, H/PD, the number of hospitalizations per person-day; dividing by PD makes the hospitalization counts

more comparable to start with, commonly done with such data. We use the log transformation to help us statisfy our typical regression assumptions of INF 511. (We'll check assumptions formally in chapter 6 and will provide more explanation of such data in INF 512. For now, just take log(H/PD) as our given response.)

The data are contained in the file `hw3.df.RDS` in BbLearn. Below, I read the data into `R` and construct a plot to get you started.

```
> hw3.df<- readRDS(file="hw3.df.RDS")
> plot(log(H/PD) ~ x, data=hw3.df, pch=as.numeric(HVI),
+      xlab=expression("Daily Maximum Temperature" ~ (x[i])),
+      ylab=expression("Hospitalization Rate" ~ (log(H[i]/PD[i]))))
> legend("topleft", legend=c("high", "med", "low"), pch=c(3,2,1),
+ title="HVI Group")
```



1. The plot, above, suggests to me that the (log) hospitalization rate is a quadratic function in temperature, perhaps shifted up or down, more or less, depending on the HVI group. Thus, we will use `lm` to regress `log(H/PD)` on the covariates of temperature, temperature squared, and HVI group. We will subtract 30 from the temperature before using `lm`; please continue reading this item before using `lm`.

   Note that the HVI variable is a factor. For this factor, by default, `R` will create 2 binary ('0/1') covariates, x_1 and x_2, one that indicates (with a 1) the medium HVI group, and

the other that indicates the high HVI group. The low HVI group is 'indicated' when both x_1 and x_2 are zero, of course. The parameters associated with the HVI factor's indicator variables, x_1 and x_2, provide for changes in the intercept for the medium and high groups, respectively, relative to the low group, whose intercept is given by the usual β_0 alone. (This 'shifting intercept among HVI groups of a quadratic relationship can be seen on the plot. Yes?) You do not have to do anything special to get `R` to code HVI groups this way.

Do not create a (log) rate variable; use the left-hand-side of the formula argument in `lm` to compute it (as in the code used to create the introductory plot).

Also, in your regression formula, subtract 30 from temperature (`x`), e.g., `I(x-30) + I((x-30)^2)`. (Beware the 'as is' function, `base::I`, which I use in our notes and mentioned in class.) This will make 30 degrees a baseline or reference temperature, i.e., when x is 30, then $(x - 30)$ in the model will be zero, and the intercept, β_0, will have the sensible interpretation of mean (log) hospitalization rate for the low HVI group. Again, do not create new variables to do this; use the formula.

Finally, do not create the indicator variables; `R` will recognize the HVI variable as a factor and add the indicator variables to the regression matrix, $\mathbf{X}$, automatically.

Show your code and output, including the usual summary of your `lm` object. (5 points)

2. What proportion of variability of (log) hospitalization rate is accounted for by the linear association with temperature and HVI group? (Linear in the parameters.) (2 points)

3. Is it appropriate to infer about the intercept, β_0, for the model considered in the previous items? Remember, we've changed the interpretation of the intercept by subtracting 30 from temperature. Thus, in the model, $(x - 30)$ is zero when temperature, x, is 30. The introductory plot may help to answer this question. Explain concisely. (5 points)

4. Perform a test to drop the HVI factor from the model. For this, you may use either the Full vs. Reduced (RSS) model approach or the GLH $\boldsymbol{C\beta}$ approach, that latter using `gmodels::glh.test`. You may have to download and install the `gmodels` package. Be sure to **(i)** state the null and alternative hypotheses in terms of the parameter symbols, as we have done in class; **(ii)** report the F statistic of the test; **(iii)** report the p-value; and **(iv)** give a brief conclusion in the context of the problem. Assume a type I error probability (significance level) $\alpha = 0.05$. Recall that the HVI factor is coded as 2 indicator columns in $\mathbf{X}$, and be sure to use numbered subscripts in your notation corresponding to the occurrence of these variables' parameters in your summary printout in a previous item. (6 points)

5. **(i)** How many parameter are in the full model (not including the error variance)? **(ii)** How many are in the reduced model? **(iii)** Give the residual (or error) degrees of freedom. **(iv)** Use the `qf` function to compute the p-value fo the previous test. (4 points)

6. We have reason to believe, a priori, that the high HVI group has a higher (log) hospitalization rate than the low HVI group. Accordingly, using the full model fit in a previous item, construct a 95% **one-sided** lower bound for the difference in the mean (log) hospitalization rates between the high and low HVI groups for the same temperature. (We want to say how large this difference is by saying it's at least as large as a lower bound with 95% confidence.) Report a brief statement along with your lower bound. Be sure to show reasonable steps towards your answer to permit partial credit. (Hint: We are talking about inferring a difference of means; recall homework 2 problem 2(a-b).) (6 points)

7. As in the previous item, we have an a priori reason to believe that the high HVI group (log) rate is higher than that of the low HVI group, at the same temperature. Conduct the one-sided test of the null hypothesis of no difference between the mean rates of these two group verses the alternative that the HVI group mean rate is higher. Use $\alpha = 0.05$. Again, give reasonable steps towards your answer to permit partial credit (state your hypotheses using parameter symbols as used in class and give the test statistic, the p-value and a statement of your conclusion). (6 points)