Math 505 Lab 4

1. (a) Write a function called `ftest` that accepts two models as inputs and returns the p-value from an F-test comparing the two models. You may use the commands `deviance`, `length(coef())`, and `length(residuals())`, which should come in handy.

   (b) Test this function using the data from Lab 2, by comparing the model `Y~X1+X2+X3` to the model `Y~1`. You should get the F-test p-value reported in the model summary for the model `Y~X1+X2+X3`. This will also happen anytime you compare a model with an intercept to the intercept-only submodel.

2. The data set `cows.txt` contains milk production values for 300 (hypothetical) cows, 100 from the Andrews Farm, 100 from the Bailey farm, and 100 from the Carter farm. Let `Y` be the first column, which contains the milk production values, and let `Farm` be the second column, which indicates the farm that each cow came from. We assume that milk production is normally distributed with constant variance, and milk production for different cows is statistically independent. The average milk production at these farms could be different though.

   (a) Use R to fit the model `Y~Farm`. Note that it fits the model

$$Y_i = \beta_1 + \beta_2 \text{FarmBailey}_i + \beta_3 \text{FarmCarter}_i + \epsilon_i, \text{ for } i = 1, \ldots, 300, \text{ where}$$

$$\text{FarmBailey}_i = \begin{cases} 1, & \text{if Farm}_i = \text{Bailey} \\ 0, & \text{otherwise} \end{cases}$$

$$\text{FarmCarter}_i = \begin{cases} 1, & \text{if Farm}_i = \text{Carter} \\ 0, & \text{otherwise.} \end{cases}$$

   (b) What are $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$? What are the average milk production estimates for each farm? (See endnotes for hints[1].)

   (c) What is the estimate for the standard deviation of milk production?

   (d) What is the dimension of the parameter vector for this model?

   (e) What is the residual sum of squares for this model?

   (f) Define the dummy variables `FarmBailey` and `FarmCarter` yourself, and verify that you get the same model by fitting `Y~FarmBailey+FarmCarter`.

   (g) Let H_0 be the hypothesis that the average milk production on the three farms is equal. State the hypothesis H_0 in terms of the β_j's.

   (h) Fit the regression model under the assumption that H_0 is true.[2] What is the average milk production in this scenario?

   (i) Use an F-test to test H_0 against the alternative that at least two farms differ in average milk production.[3]

3. Now, fit the following (equivalent) model that we looked at in class

$$Y_i = \mu_A \text{FarmAndrews}_i + \mu_B \text{FarmBailey}_i + \mu_C \text{FarmCarter}_i + \epsilon_i, \text{ for } i = 1, \ldots, 300.$$

You will need to create an additional dummy variable `FarmAndrews`. Also, this model does not have an intercept, so you can fit it with the formula

`Y ~ FarmAndrews+FarmBailey+FarmCarter-1`.

(a) What are $\hat{\mu}_A$, $\hat{\mu}_B$, and $\hat{\mu}_C$? What are the average milk production estimates for each farm?

(b) What is the estimate for the standard deviation of milk production?

(c) What is the dimension of the parameter vector for this model?

(d) What is the residual sum of squares for this model?

(e) Again, let H_0 be the hypothesis that the average milk production on the three farms is equal. State the hypothesis H_0 in terms of the μ_j's.

(f) Use an F-test to test H_0 against the alternative that at least two farms differ in average milk production, using the model from this problem.[4]

(g) Do your results on this problem agree with those from problem 2?

4. The file `SATMotherEd.txt` contains the math SAT scores (`Y`) and mother's education level (`mother`) for 3146 students.

(a) Verify that there are 3146 rows of data. Create a histogram for `Y` and explore mother's education level using the command `table(mother)`.

(b) Since `mother` is a factor variable (categorical variable), we can list its levels (possible values) using the command `levels(mother)`. This command simply lists the levels in alphabetical order.

(c) Now, fit a linear model called `model` for predicting math SAT score from mother's education level. Note that R creates dummy variables for the levels "Bachelor Degree"..."Some High School". The only level without a dummy variable is the first one, "Associate/two-year degree".

"Associate degree" is the reference level for the model, and all coefficients represent comparisons to this level. For instance, the coefficient on `motherBachelor Degree` of -4.288 indicates that students whose mothers had bachelor degrees had math SAT scores 4.288 points lower on average than students whose mothers had associate degrees. Also, the intercept of 508.571 is the average math SAT score for the associate degree level.

It's not very helpful to make comparisons to the associate degree level, since only 35 students were in this category. It is often preferred to have the most common category, in this case "Some College", be the reference level. To do this, use the command `mother=relevel(mother,"Some College")`. After releveling the `mother` variable, refit the model.

(d) Using the model summary, find the average math SAT score for students in each of the eight categories. Which categories have students with SAT scores that are statistically significantly different from the "Some College" category?

(e) There are some problems with this model. Many of the levels are not statistically significant, and the number of parameters (8), is high, considering we're really only dealing with one input variable. Perhaps we can join similar categories together, reducing the number of categories and the number of parameters. Intuitively, these categories seem to be ordered:

No High School, Some High School, High School Diploma, Some College, Associate Degree, Bachelor Degree, Graduate Degree

"Not available" doesn't really fit into this ordering, so we will keep it separate. This suggests combining categories as follows:

$$\begin{pmatrix} \text{No High School} \\ \text{Some High School} \end{pmatrix} \to \text{None}$$
$$\begin{pmatrix} \text{High School Diploma} \\ \text{Some College} \end{pmatrix} \to \text{HS Diploma}$$
$$\begin{pmatrix} \text{Associate Degree} \\ \text{Bachelor Degree} \\ \text{Graduate Degree} \end{pmatrix} \to \text{Degree}$$
$$(\text{Not Available}) \to \text{Not Available}$$

Also, the coefficients within each of these bigger categories are relatively similar, for instance, -22.8 and -19.5 for the "None" group. This suggests that combining groups in this way might work well. Let's see! To do this in R, use these commands

- `motherrecode=c("HS","Degree","Degree","Degree","HS","None", "Not Avail","None")`
- `mother2=motherrecode[mother]` (`mother2` is the `mother` variable with categories combined.)
- `mother2=as.factor(mother2)` (This forces `mother2` to be a factor variable.)
- `mother2=relevel(mother2,"HS")` (This makes the most common level, "HS", the reference level.)

Verify that these commands actually combine groups as desired using the `table` command.

5. Fit a model called `model2` that predicts math SAT based on `mother2`. Comment on any features of this model that seem interesting to you.

6. `model` can be specified by the regression equation

$$Y_i = \beta_1 + \beta_2 \text{motherAssociate degree}_i + \cdots + \beta_8 \text{motherSome High School}_i + \epsilon_i.$$

Show that `model2` is equivalent to the statement $\beta \in V_0$, for some subspace $V_0 \leq \mathbb{R}^8$, that is, show that `model2` is equivalent to a certain system of linear equations involving the β_j's. Therefore, we can perform the F-test

$$\text{H}_0 : \beta \in V_0 \text{ vs. H} : \beta \in \mathbb{R}^8, \text{ which is equivalent to}$$

$$\text{H}_0 : \texttt{model2} \text{ is valid vs. H} : \texttt{model} \text{ is valid.}$$

In other words, `model2` is a *submodel* of `model`, and we can compare them with an F-test.

7. Compare `model2` and `model` with an F-test. Based on the results, is it reasonable to combine levels of the `mother` variable in this way?

8. If it's reasonable to combine these categories, is it desirable? (Consider the number of parameters for each model.)

9. Can you improve the model even more by doing something with the "Not Available" category? If so, fit the improved model and check its adequacy by comparing it to `model` (technically, we need to check the adequacy of `model` first, but you can ignore this issue for now).

Notes

[1]The average milk production and the $\hat{\beta}_j$'s are not the same.

[2]If H_0 is true, then $\beta_2 = \beta_3 = 0$, and the model is $Y_i = \beta_1 + \epsilon_i$, which is the intercept-only model.

[3]H_0 is the intercept-only model, and the alternative hypothesis H is the model `Y~Farm`.

[4]If H_0 is true, then $\mu_A = \mu_B = \mu_C$. Let's call the common mean μ_0. Then the model is

$$\begin{aligned} Y_i &= \mu_A \text{FarmAndrews}_i + \mu_B \text{FarmBailey}_i + \mu_C \text{FarmCarter}_i + \epsilon_i \\ &= \mu_0(\text{FarmAndrews}_i + \text{FarmBailey}_i + \text{FarmCarter}_i) + \epsilon_i \\ &= \mu_0 \cdot 1 + \epsilon_i, \end{aligned}$$

the intercept-only model. The alternative hypothesis H is given by

`Y ~ FarmAndrews+FarmBailey+FarmCarter-1`.