STAT 422/722 Spring 2016 Homework #2
Limited Solutions

Professor Adam Kapelner

Wednesday 22nd February, 2017

Problem 1

We will be investigating equivalence testing.

(a) [easy] In the context of linear or logistic regression, if you want to prove that a predictor has a linear effect on the response (controlling for other variables), what are the null and alternative hypotheses?

$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0$$

(b) [harder] In the context of linear or logistic regression, if you want to prove that a predictor does *not* have a linear effect on the response (controlling for other variables), what are the null and alternative hypotheses?

$$H_0 : \beta_j \neq 0$$
$$H_a : \beta_j = 0$$

The null and alternative are "flipped" relative to (a).

(c) [easy] You collect four data points

predictor	response
2.47	0.50
0.57	1.95
0.84	1.91
2.18	2.51

Test the theory in (a)

Assume $\alpha = 5\%$ for testing purposes. Note: all code can be found in the file `sol.R` in this directory. Output is equivalent in JMP. `R` gives us:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2698     1.0062   2.256    0.153
x            -0.3646     0.5838  -0.624    0.596
```

Since the p-value is $0.596 \not< 0.05$ this means we fail to reject H_0.

(d) [harder] Test the theory in (b). Use $\delta = 0.5$ as a margin of practical equivalence

Assume $\alpha = 5\%$ for testing purposes. We use the same output as above. We note that a 95% CI for β is $[-.3646 \pm 2 \times 0.5838] = [-1.532, 0.803]$. Since this is not a proper subset of (i.e. within) $[-\delta, +\delta] = [-0.5, 0.5]$, we fail to reject H_0.

(e) [difficult] How can you get both the answer to (c) and the answer to (d) at the same time? Discuss.

Problem 2

We will be investigating dredging and multiple testing corrections. I have provided a data file for you called "xyrand.csv" located here (right click and downlod from the browser). This file is fully random data from a standard normal and thus there is no systematic connection between the column y and any of the x_j columns.

(a) [easy] Run a regression of y on the x_j's and report R^2. Why is this R^2 not *exactly* zero?

(b) [easy] Which variables were significant and what are their significance levels? Why should any variables be significant in the first place if they're all just $\overset{iid}{\sim}$ random realizations?

(c) [harder] Calculate the probability you see this many significant variables *or more* in a 50-predictor linear regression. Is your number of significant variables "expected"?

(d) [easy] Calculate both a Sidak and a Bonferroni corrected individual α that preserves 5% familywise error.

(e) [easy] Using the Sidak and/or the Bonferroni correction, are there any significant variables anymore? Yes/no

(f) [harder] Explain precisely what I would need to simulated in this same setup with one response and 50 predictors randomly realized to expect one significant variable if the familywise correction is employed.

(g) [harder] Report the overall F value and it's corresponding significance level. Explain how it is possible that there exist t tests which are significant for some linear predictors but the F test is not.

Problem 3

These are some conceptual questions concerning hypothesis testing, Type I errors, Type II errors and power.

(a) [easy] Given some predetermined level of α we have two ways of setting up hypothesis tests:

$$H_0 : \text{UFOs do not exist}$$
$$H_a : \text{UFOs do exist}$$

and the inverse:

$$H_0 : \text{UFOs do exist}$$
$$H_a : \text{UFOs do not exist}$$

Which set of hypotheses should be employed and *precisely* why?

(b) [harder] You cannot convince your friend. In the hypothesis testing framework there are two separate reasons why he could remain not convinced. What are they?

(c) [difficult] The Shapiro-Wilk Test of Normality is used to assess normality of a given sample of data. It is a goodness-of-fit test where

$$H_0 : \text{ the data generating process is normal}$$
$$H_a : \text{ the data generating process is not normal.}$$

Usually, you want to prove normality (e.g. the case of testing the residuals from a linear regression). Why does this test reward small sample sizes?

Problem 4

These questions will be about extrapolation and generalization.

(a) [harder] Is model extrapolation and model generalizablity the same concept (lecture 3, slide 6)? Discuss.

(b) [difficult] Provide an example of a model that you use regularly that does not generalize to the observations you use to predict with it.

(c) [harder] Run lines 5–27 of the lecture 3 R demos. Which of the three models would be the worst "extrapolator" and why?

(d) [difficult] You are provided a new $\boldsymbol{x}^*$ with p features in which are to guess y. How would assess extrapolation? Explain.

Problem 5

These questions will be about optimal design.

(a) [easy] In the case of a simple linear regression with $n = 20$ points where x ranges from 6 to 17, what would be the optimal design?

(b) [harder] Show that for fixed n under least squares regression that the optimal design is half the points on the minimum and half the points on the maximum.

(c) [easy] Take the case of $n = 20$ and three continuous predictors ranging in $[0, 1]$. Use JMP to create an optimal design for the model with all factorial interactions and all polynomials up until degree 3. Take the optimal design right click and make data table. Then sort the data table first by x_1 then by x_2 and by x_3. Are they all about the same? Yes/no.

(d) [difficult] The optimal design in the previous problem is what you'd probably like to do for non-parametric linear model with lots of interactions and curves. What is the takehome message in this case?

(e) [E.C.] Your goals are prediction and you have the choice bertween D-optimality and I-optimality. Which is likely better and why?

Problem 6

These questions will be about logistic regression using the Telecom Churn dataset that can be downloaded here.

(a) [easy] Give an expression for the conditional mean in a logistic regression problem with p features using the standard logistic regression assumptions.

(b) [easy] Do the multivariable logistic regression in class with target response churn (remember to delete those 6 variables which are fully collinear). Provide an interpretation on the monthly charges coefficient.

(c) [E.C.] If we used the `cloglog` link function and got the same coefficient, what would be the interpretation?

(d) [easy] Predict the mean probability of churn for a senior citzen who has been with the company for 36 months and then predict whether or not this person will churn.

(e) [E.C.] In my regression output, the χ^2 value for `senior citizen` is 14.37. Calculate the standard error from this.

(f) [harder] Test whether or not removing *both* gender and partner makes a difference in terms of linear predictive power versus the full model from (b). You will need a table of critical values of the χ^2 distribution at $\alpha = 5\%$. See below.

p	χ^2_p critical value
1	3.84
2	5.99
3	7.81
4	9.49
5	11.07
6	12.59
7	14.07
8	15.51
9	16.92
10	18.31
11	19.68
12	21.03
13	22.36
14	23.68
15	25.00
16	26.30
17	27.59
18	28.87
19	30.14
20	31.41

(g) [easy] Use the model from (b) and use JMP to compute the AUC and misclassification error.

(h) [difficult] Attach a graph of the false negative proportion versus the false positive proportion. Is this more useful than an ROC curve for the case of churn?

(i) [E.C.] Create a detection error tradeoff plot for this dataset (no need to use normal deviates for x and y axes).

(j) [harder] Imagine the cost ratio is 7.5:1 for the more costly mistake. What is the p_0 of the optimal model?

(k) [easy] Why is this p_0 less than the naive value of 50%?

(l) [harder] Create a flowchart illustration of the optimal model in (j) similar to lecture 3, slide 35.

Problem 7

These questions will be about survival regression using the NetLixx dataset that can be downloaded here. The response is the "time" variable and it's measured in days (ignore the "start" and "followtime" variables. The churn variable is 1 if there is churn.

(a) [easy] As we saw in class, the exponential model is not a great real-world model. Nevertheless, assume the model assumptions and run a standard exponetial model with all three predictors. Report overall model fit, each variable's estimate and their significance levels.

(b) [easy] Should dredging be considered a problem here when we're looking at all these $H_0 : \beta_j = 0$ tests? Yes/no and why.

(c) [easy] Interpret the value of the fitted coefficient for coupon.

(d) [easy] Predict the mean survival time for a new female of age 30 without a coupon.

(e) [E.C.] Calculate a 95% CI for mean survival time for the new person in (d).

(f) [E.C.] Calculate the in-sample R^2 for the non-censored rows. How does the model do?

(g) [E.C.] Fit a Weibull model to the same data with the same response, censoring and predictors. Does it fit better? Why / why not?

Problem 8

These questions will be about our pivot from parametric linear models to non-parametric non-linear models. To illustrate, we will use the white wine dataset that can be downloaded here.

(a) [harder] Let's do an exploratory data analysis. Do Fit y by x in JMP where y is quality and x are all variables. Then, fit polynomials of degree 2 to each of them. Which ones have significant squared terms at $\alpha = 5\%$? Make sure you do a multiple testing correction.

(b) [harder] Let's now look at all first-order interactions. These are interactions that are between two variables. Use the fit model, highlight all variables and do macros... factorial to degree (the default degree in JMP is two indicating first-order interactions). Which interactions are significant at $\alpha = 5\%$? Make sure you do a multiple testing correction.

(c) [easy] Interpret the coefficient on `density` $\times$ `alcohol`.

(d) [easy] Based on your answers to (a) and (b) would it be fair to say that the true model is non-linear? Yes/no.

(e) [easy] Based on your answers to (a) and (b) would it be fair to say that you get better predictive power if you add in some quadratic terms and interactions? Yes/no.

(f) [easy] If you add some quadratic terms and interactions and interactions, what are the three things you are giving up in your model?

(g) [easy] The vanilla multivariable regression gives $R^2 = 28\%$. Now fit a model with *all* interactions. Use "full factorial" in the options. What is R^2 in this new absurd model? What are the degrees of freedom?

(h) [E.C.] What are the significant variables now? You may have to use a t-calculator.

(i) [harder] Prove that you overfitted.

Problem 9

We will now explore the concept of overfitting.

(a) [easy] What are you fitting when you "overfit" and why is that not a good idea?

(b) [difficult] What are you underfitting when you "underfit" and for what purpose is this usually done?

(c) [harder] Explain why running a linear regression with $n = p + 1$ (where the $+1$ is for β_0, the intercept) guarantees you overfit always. Note: we are assuming each of the p features are not perfectly collinear with any of the others.

(d) [difficult] The illustration in lecture 4, slide 32 shows the final model $\hat{f}$ being built from the entire dataset. Why would you do this?

(e) [easy] Explain why using the test set for more than one model generalization error estimate causes you to overfit.

(f) [difficult] Describe a scenario where you would want to use out-of-sample validation with a test set consisting of 90% of the entire dataframe.