# STAT 422/722 Spring 2016 Homework #2 Limited Solutions

Professor Adam Kapelner

Saturday 25th 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 $\beta$ 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?

$R^2 = 29.2\%$ and it's not zero due to chance capitalization of $x$'s being randomly related to $y$.

(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?

$x_{26}$ with a $p$-value of .0499, $x_{27}$ with a $p$-value of .0491, $x_{33}$ with a $p$-value of .0448 and $x_{38}$ with a $p$-value of .0320. These four variables should not be significant in the first place since there is no real correlation. They are only "significant" due to chance.

(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"?

This is a binomial exercise. You have a probability of falsely rejecting $H_0$ if $H_0$ is true (our case of no linear correlation) of $\alpha = 5\%$. Thus we have:

$$\mathbb{P}\left(N \geq 4\right) = \sum_{i=4}^{50} (5\%)^i \, (95\%)^{50-i} = 1 - F(3) \approx 24.0\%$$

This type of calculation will not be on the exam.

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

Sidak is not covered on the exam. The Bonferroni $\alpha$ is merely $5\%/50 = 0.01\%$.

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

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.

The Bonferroni now controls familywise error rate. The "family" here is the set of the 50 $t$ tests. Thus it guarantees that getting one or more Type I errors in that family is at most $\alpha = 5\%$. However, if I run the entire family 20 times. That's 20 simulations of 50 $t$-tests each, I will expect one Type I error since $20 \times 5\% = 1$, the expectation calculation.

(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.

$F = 1.2285$ which has a significance level of 17.34% i.e. not significant at $\alpha = 5\%$. The $F$ test is testing overall usefulness of the model, not individual variables. The $F$ test "knows" that regressions with many, many variables will chance capitalize and that degree of explantory power is expected. It only allows models to pass which are above the expected level of chance capitalization — those are "significant" models.

## 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 $\alpha$ 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?

The first set is employed. $H_a$ is the theory you wish to prove. So you are intellectually honest and assume your theory is *not* correct a priori; this is $H_0$. Then, you let the data speak for itself. If the data convinces you overwhelmingly (where "overwhelmingly" is the level of skepticism defined by your level $\alpha$), then you can accept your theory, but only then!

Contrast that to the second set. Here, you have begun with the belief that your theory is correct. Your theory will remain correct unless the data convinces you otherwise. This is not an intellectually honest means of reasoning.

(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?

1. His $\alpha$ is too low i.e. he is too skeptical. No amount of data will convince him; he will never budge from his $H_0$.
2. You didn't provide enough data — you are "underpowered".

Parenthetically, these two reasons are not either-or. In practice, it may be a combination of both. When statisticians practice, we set up $\alpha$ before hand and we agreed as a community it should be 5% or 1%. Conditional on accepting this community standard, the reason then would be #2.

(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?

Reference (a). You've begun by believing what you wanted to prove! So not having data is a good thing — less chance you challenge your beliefs!

## 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.

Yes. Models generalize to new observations that are within the measurement / covariate space of those observations previously seen i.e. the rows in the historical dataframe you made use of when you constructed the model. Extrapolation is predicting outside of this range. It is indeed failure to generalize.

(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.

Prediction models for stock returns. Non-stationarity (where $f$ is time-dependent) is another failure to generalize.

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

Polynomial regression. The $\hat{f}$ is too erratic outside of the region / range of the historical data. It is less likely that the true $f$ conditional expectation function behaves like this outside of the range. Better extrapolators are linear regression and decision trees.

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

Check each of the $p$ measurements and see if they are outside of the ranges of those in the historical dataframe. That's a first pass at a solution. But this is a deep problem since it's multidimensional.

## 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?

10 observations at $x = 6$ and 10 observations at $x = 17$.

(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.

The equation for the variance of $\hat{\beta}$ in one dimension is a fraction with $\sum(x_i - \bar{x})^2$ in the denominator. The way to minimize the fraction is to maximize the denominator. This is done by choosing half of the points at the minimum and half at 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.

Yes.

(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.

$$\hat{p} = \hat{p}(\boldsymbol{x}^*) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\ldots+\hat{\beta}_p x_p}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\ldots+\hat{\beta}_p x_p}}$$

(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.

When running the logistic regression, we get $\hat{\beta} = 0.0217$. This means that if monthly charges increase by $1, the log-odds of churning increase by 0.0217 keeping all other variables constant in a naturally observed customer. Note that this interpretation may not be valid if the monthly charges variable has high collinearity with other variables in the regression.

(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 citizen who has been with the company for 36 months and then predict whether or not this person will churn.

We run another regression now with two variables: senior citizen and tenure and the result using the answer for (a) above is:

$$\hat{p} = \hat{p}(\boldsymbol{x}^*) = \frac{e^{-0.123+1.047x_1+-0.0405x_2}}{1+e^{-0.123+1.047x_1+-0.0405_p x_2}}$$

We then substitute 1 for $x_1$ (senior citizen) and 30 for $x_2$ (tenure) and compute:

$$\hat{p} = \hat{p}(\boldsymbol{x}^*) = \frac{e^{-0.123+1.047+-0.0405(30)}}{1+e^{-0.123+1.047+-0.0405(30)}} = \frac{e^{-0.291}}{1+e^{-0.291}} = \frac{0.748}{1+0.749} = 0.428$$

which is the predicted probability of churning. Using the naive classification rule of $p_0 = 50\%$ we predict the $\hat{\boldsymbol{y}} = 0$ i.e. this person is not predicted to churn.

(e) [E.C.] In my regression output, the $\chi^2$ 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 $\chi^2$ distribution at $\alpha = 5\%$. See below.

| $p$ | $\chi^2_p$ 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 |

The log likelihood for the full model is -2974.327 and the log likelihood for the reduced model is -2974.343. Thus the chi-squared test statistic is $Q = 2(-2974.327 - -2974.343) = 2(0.016) = 0.032$. The critical value for this test is 5.99 (it's the value with df $= 2$) and since 0.032 is less than 5.99 we fail to reject.

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

We get AUC $= 0.84037$ and we get misclassification error from computing directly from the confusion matrix $\frac{887+537}{7032} = 0.203$.

(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?

Here, we need to export the ROC table as a dataframe in JMP then make an extra column called "cost" with the formula $7.5 \times FN + FP$ then we sort by this new coumn and the first row's $p_0 = .1512$.

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

In this asymmetric cost scenario, false negatives are much more costly than false positives. Thus, it makes sense to be conservative in what we call a negative. The way to do this is to reduce $p_0$ to obtain fewer predicted negatives.

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

The answer really is on the slide.

## 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.

The model is significant: $\chi^2_3 = 64.6$ with a significance level of less than 1 in 10,000. Female is significant with $\hat{\beta} = 0.089$ and a significance of 0.0165 and age is significant with $\hat{\beta} = 0.0059$ and a significance level of less than 1 in 10,000.

(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.

Yes. You are running three tests.

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

With all other variables remaining constant, adding the coupon would account for tenure to be multiplied by $e^{-0.01249} = 0.989$ for a naturally observed customer.

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

$$\hat{y} = e^{\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \ldots + \hat{\beta}_p x_p^*} = e^{6.466+0.0894(1)+0.00559(30)+-0.0125} = e^{6.7106} \approx 821 \text{ days}$$

(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.

The multiple testing correction for 11 quadratic tests yields $\alpha = 0.05/11 = 0.00455$.

At this level, the significant squared terms are fixed acidity, volatile acidity, citric acid, residual sugar, free sulfur dioxide, total sulfur dioxide, density, alcohol.

(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.

The multiple testing correction for 55 interaction tests yields $\alpha = 0.05/55 = 0.00091$.

At this level, the significant squared terms are volatile acid × alcohol, free sulfur dioxide × total sulfur dioxide and free sulfur dioxide × sulphates.

(c) [easy] Interpret the coefficient on `density` × `alcohol`.

Even though this coefficient is not significant, we will interpret it. There are many ways to view this:

(a) If density increases by one unit, the slope of the alcohol variable will move by -23.18.

(b) If density increases by one unit, the response will move by -23.18 plus the coefficient on density which is -160.37.

(c) If alcohol increases by one unit, the slope of the density variable will decrease by 23.18.

(d) If alcohol increases by one unit, the response will change by -23.18 plus the coefficient on alcohol which is 0.11.

Always "controlling for everything else" plus the usual other caveats.

(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.

Yes.

(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.

Yes.

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

We give up simplicity, interpretability and inference (on individual effects).

(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?

$R^2 \approx 75.1\%$ and df $= 2036$.

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

(i) [harder] Prove that you overfitted.

You can do a $K$-fold CV in JMP by fitting the same model using stepwise, enter all and then click the $K$-fold option in the main menu. It takes awhile to compute but the $R^2$ oos on 5-folds is -1000000 which is not only lower than the in-sample $R^2$ of 75% but it is a complete disaster!

## 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?

You are fitting the irreducible noise term $\mathcal{E}$ which is defined to be independent of your predictors $x_1, \ldots, x_p$. This is bad since you are effectively fabricating a model. Fabricated models are not data-driven and may not generalize well.

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

You are underfitting $f$, the true conditional expectation function of the predictors in your dataframe $x_1, \ldots, x_p$. You usually purposely underfit because you make parametric assumptions e.g. the linear model. This buys you a lot though — inference, simplicity and interpretability.

(c) [harder] Explain why running a linear regression with $n = p + 1$ (where the $+1$ is for $\beta_0$, the intercept) guarantees you overfit always. Note: we are assuming each of the $p$ features are not perfectly collinear with any of the others.

When the number of degrees of freedom equal the number of data points, the company can minimize SSE by fitting a hyperplane that passes through all $y$ values. Thus SSE is 0 and $R^2 = 100\%$. Since in any modeling problem in the real world, there is by definition the irreducible error $\mathcal{E}$, $R^2$ cannot ever be 100%. Thus, you have unequivocally overfit the data.

(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.

It doesn't "cause" you to overfit, but it opens the door to it. You may obtain a bad model (one that doesn't generalize well) and by sheer luck have it shine on the test set (a la the spurious correlations we've seen). The more models you check, the more likely this is to happen.

(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.