Stat 305A: Linear Models **Autumn 2019/20**

Lecture 19: Summary

Lecturer: Art B. Owen *December 5*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. For instance, citations are mostly omitted or imprecisely made. The notes are meant as a memory aid for students who took stat 305A at Stanford University. They appear in* `http://statweb.stanford.edu/~owen/courses/305a/` *where you may find the most recent version. Stanford University holds the copyright.*

This lecture was a summary of the course, a description of good followup courses, and some thoughts on the applied statistics enterprise in general.

19.1 Recap

This course is about applied statistics and also about the linear model, so it uses the linear model to illustrate some more widely applicable points. We have (X_i, Y_i) pairs and we want to model how Y depends on X. One very concrete way to describe that is we would like to be able to predict say Y_{n+1} given that $X_{n+1} = x_{n+1}$. We want a bit more too. We have a persistent interest in whether what we are doing generalizes. If the data are IID then we can generalize to the distribution they've come from but sometimes we want more.

We begin by turning each $X_i \in \mathbb{R}^d$ into some features $Z_i \in \mathbb{R}^p$. We looked at lots of different ways to do that depending on what X's we are given (none, and on what goals we have and on what we believe to be important.

The most basic model was that $Y \sim \mathcal{N}(Z\beta, \sigma^2 I)$ or sometimes just $Y \sim (Z\beta, \sigma^2 I)$. We need normality to derive exact p-values and confidence intervals. We don't need it for Gauss-Markov or for asymptotic p-values and confidence intervals on β. We do need it for confidence intervals on σ and prediction intervals for Y_{n+1}. Too bad that this is complicated. It is what it is.

Using calculus, linear algebra, geometry and then the singular value decomposition for computation, we can get the least squares estimate $\hat\beta$. Sometimes it is unique; sometimes not.

Using distribution theory we get confidence intervals, p-values and hypothesis tests out of these distributions: t, F, χ^2, noncentral χ^2, F and noncentral F. There is even a doubly noncentral F though it comes up less often. From those distributions we can also work out power and pick sample sizes.

What we got was $\hat\beta \sim \mathcal{N}(\beta, (Z^\mathsf{T}Z)^{-1}\sigma^2)$ and $s^2 \sim \sigma^2\chi^2_{(n-p)}/(n-p)$ under normality. In a way, these two things are most of what we need to know. In another way, the consequences of these results are not obvious until you think of specific use cases. For instance with plain linear regression of Y on a scalar x it emerges from working things out that it is better to have a large sample variance of x_i in the data. Thinking further it is clear that a nonsignificant slope estimate could just be because the x_i we had were only varied over a narrow range. That could happen if our process controls and selects only x_i in a narrow range (examples are selection of people into school or jobs, as well as industrial controls of some variable for safety).

The matrix $Z^\mathsf{T}Z$ and the model $Y \sim (Z\beta, \sigma^2 I)$ come in the middle of a pipeline, after some domain specific reasons have chosen our data and before some domain specific consequences will follow from our choices. In applied settings we have to think about several parts of the pipeline. The interface between statistical

© Stanford University 2019

reasoning about a model and the domain science is not clean and clear. It is a tangle. Calculus is very different. One can much more easily figure out how it works and how to use it. Nobody differentiates when the should have integrated. Things like that happen with statistics.

19.2 Problems and fixes

We looked at what can go wrong with the linear model. Maybe $\varepsilon = Y - Z\beta$ is not Gaussian. Of course it is not Gaussian. How could that ever happen? It could be close to Gaussian, or it might not be. The central limit theorem applies to $\hat{\beta}$ so even if the errors are not Gaussian, the exact answer we derive under the Gaussian assumption can be ok. Also, $s \to \sigma$ by the law of large numbers, and for large degrees of freedom our t statistics work out ok. So then do p-values, hypothesis tests and confidence intervals (collectively: inferences) on $c^{\mathsf{T}}\beta$.

If the variance is not constant over observations, then $\hat{\beta}$ is still unbiased but there is a better unbiased linear estimate. The most serious issue is that our estimate of $\mathrm{var}(\hat{\beta})$ depends strongly on equal variances for the errors. Large n will not fix the problem and then we will have incorrectly calibrated inferences. One fix is to somehow model the unequal variances and use generalized least squares (GLS). Or, we could bootstrap (x, y) pairs. Or use the wild bootstrap if we want to fix the x_i. Bootstrapping residuals will just give approximately the incorrect inferences based on equal variances. [Play sad trombone sound here.]

If the ε_i are dependent, then the situation is much more difficult. We might be able to model the dependence, estimate parameters in that model, and then use the estimated variance in GLS. Our model won't be perfect, but we can at least expect improved inferences. If the dependence is due to a hierarchical structure (e.g., beginning readers within classrooms and/or classrooms within schools) then it is approximated by a random effects ANOVA model. Arbitrary unknown dependence will spoil inferences.

Dependence can even interfere with holdout methods like cross-validation that are based on IID sampling. The held-out data can be correlated with the held-in data. Cross-validation can be motivated by supposing that the data we have are IID from some population. If we randomly downsample then the held in and held out data are each IID samples of some smaller size from that population. If the data were not IID in the first place, then the story behind cross-validation is much more tricky. I think that this is an area that could use some additional research.

Bias is an extremely hard thing to handle. We write $\varepsilon_i = Y_i - Z_i^{\mathsf{T}}\beta$. Simply using symbols like e_i and ε_i and ϵ_i puts us in mind that they should have mean zero and be unrelated to Z_i. However, our choice of symbol does not change the world. The functional form of $\mathbb{E}(Y \,|\, X = x)$ might not be our $z(x)^{\mathsf{T}}\beta$. If n is large, then we can fit ever more flexible models and capture more of what is important in the conditional mean. What if some variable or variables are left out? Then our model lumps them into ε. The β we estimate without those variables $W_i \in \mathbb{R}^r$ will not match what we would estimate with them. If there really was a causal link from $(X_i, W_i) \in \mathbb{R}^{d+r}$ to Y we should not expect to correctly estimate it from a regression of Y_i on Z_i constructed solely from X_i.

If all we want is a way to predict Y_i from X_i, and we have no plans to make any changes to what we do based on that model, and we are not making causal interpretations, then we are less affected by missing variable bias. We might get less accurate predictions, but they might still be ok. There is at least a small role for being able to predict well without understanding why it works and without planning to take any actions. Even there, having some understanding and interpretation can guide us in understanding circumstances where those predictions might stop working in the future.

Outliers can be viewed as an extreme version of non-normality or as data that somehow come from a different process than the one most of the data follow. An outlier Y_i value that coincides with an X_i having large hat

matrix diagonal H_{ii} can do the most damage to our model. Outliers are hard to find in high dimensional settings. Multiple outliers can mask each others' presence. They can also cause 'swamping' where some good data start to look like outliers. We looked at some super robust fitting methods like least trimmed squares.

19.3 Automatic variable selections

For $Z_i \in \mathbb{R}^p$ with a very large p, there is a bias-variance tradeoff in picking which q of those p features to include in the model. Larger q brings more variance in the fitted values but less bias. We looked at AIC and BIC and GCV to choose the model. Also stepwise methods.

The methods from the prior paragraph are pretty inelegant. More principled regularizations are based on ridge regression and the lasso. We also considered regression on principal components. These methods have regularization parameters that can be tuned using a holdout sample. Yet another holdout set can be used to judge the accuracy of the resulting method. Ridge and lasso have Bayesian interpretations that help us understand what they're doing, though those are not the usual models that one turns to in a Bayesian approach. These methods bias $\hat{\beta}$ towards zero. Despite this bias, they can improve predictions by reducing variance especially when p is large.

Something to avoid at all costs is a search for a model that gives you the desired signs on a few of your favorite variables regardless of all else. If you really know those signs, then pick a Bayesian method that enforces them, and tell everybody that that is what you've done, so it does not leave them thinking that those signs came from the data.

These auto-tuned linear models can be viewed as black box predictors and their accuracy can be modeled using holdouts. It is problematic to get a confidence interval for a β_j. This is the subject of much current research. One very thorny issue is that there are 2^{p-1} different true values for whichever β_j you're thinking of, depending on which other Z_{ij} are included in the model. Which of those do you want an interval for? If your data set grows later, the selected subset is likely to change. If you always keep the intercept then there are still 2^{p-2} different true β_j's for each of the other parameters.

[There is theory to show that some algorithms will correctly identify all of the nonzero β_j and zero out all of the others as $n \to \infty$. Those however require strong assumptions such as the non-existence of nonzero β_j that are close to zero, and that the features with $\beta_j = 0$ have not got much correlation with other features that have $\beta_j \neq 0$.]

19.4 Followups

From here there are followup courses. Some look at $Y_i \in \{0, 1\}$ or $\{0, 1, 2, \dots\}$ or $\mathbb{R}^p$. Some focus on Bayesian methods. Some focus on much more computationally sophisticated algorithms than least squares. Some on how to come closer to a causal conclusion from observational data. There you can get causal conclusions though they rest on potentially untestable causal assumptions. Causal findings, if and when you can get them, have much greater value than associations.

We had two lectures introducing Bayes for the linear model. Also some coverage of causal inference topics (Neyman-Rubin view of randomized testing, regression discontinuity and instrumental variables) that are closely tied to linear modeling.

Most of the time it seems that our data come to us for free, either because they are gathered cheaply by a computer doing something else (like billing) or because somebody else has shared a data file with us. Deciding exactly how to gather that data comes up in sampling and in experimental design. We regularly

have sampling courses an sometimes experimental design.

19.5 The pipeline

Picking a model and analyzing data come with some context. They are middle steps in a big pipeline that starts with how the data arose and carries on to what we are going to do with our knowledge and how that will affect us and others.

I think it is pretty hard to anticipate what the consequences of $(Z^\mathsf{T}Z)^{-1}$ and other things are going to be without some examples to think of. There are lots of things to watch out for, and there seems to be no checklist. When helping somebody with their data you must ask some nosy questions about their goals and assumptions and what the data mean.