UNM Statistics Qualifying Exam Take-Home
August 2018

Due by 3:00pm on Monday, 13 August 2018. Email to Ana Parra Lombard at <aparra@math.unm.edu>. Make sure to identify your submission using your selected CODE WORD only; do not give your name or your UNM ID number. Do not submit a printed copy of your solutions.

Instructions:

This exam consists of two problems, given below. You will prepare a report detailing your work on both problems. Your report is to be typed, double-spaced, using no smaller than 10pt font, and with margins no smaller than 1in. Your report for each problem should be no longer than four pages. An additional four-page appendix is allowed for each problem, but this will only be examined at the discretion of the graders. (This can be a good place to include supplementary tables and figures if you run short on space. If they are well labeled and well referenced in the text, the likelihood we will examine your appendix material increases substantially.)

In your report, RAW AND UNINTERPRETED COMPUTER OUTPUT IS UNACCEPTABLE. Each figure and table should be captioned, and in the text of the report you should explain to the reader why you have chosen to highlight this table or figure. Organize the report to tell the story you uncovered, not the circuitous path you may have taken to get there. Remember that even that best data analysis is worthless if your reader cannot understand it.

You may not consult any other person when working on this exam or discuss your exam with anyone else regardless of whether or not that person is also taking the exam. You may use your course notes as well as any available books or web resources for the exam. Where possible, provide citations for methods you find outside your course notes. For clarifying questions regarding these problems, contact Dr. Fletcher Christensen, <ronald@stat.unm.edu>.

Problem 1:

The data for this problem relate to home prices in the Albuquerque metropolitan area and were collected on 22 July 2018. The dataset is available at `http://www.stat.unm.edu/~ronald/datasets/homeprices.csv`. Variables in this data set include:

- *Address* – The street address of the property being offered for sale.
- *Zip Code* – The zip code of the property being offered for sale.
- *SqFt* – The internal size, in square feet, of the property for sale.
- *Beds* – The number of bedrooms in the property for sale.
- *Baths* – The number of bathrooms in the property for sale.
- *Price* – The price being offered by the sellers of the property (list price), as of 22 July 2018.
- *Est* – An estimate of the approximate value of the property, determined by a major property-listing website.
- *Lat* – The latitude at which the property is located. Numbers closer to zero mean the property is further south.
- *Long* – The longitude at which the property is located. Numbers closer to zero mean the property is further east.

Homes for sale were randomly sampled from each zip code in the Albuquerque metropolitan area. These data constitute a sample stratified by zip code rather than a simple random sample of houses in the area. Your objective in this problem is to build a model to predict the list price (i.e. the price being offered by the sellers) for homes in the Albuquerque metropolitan area using the physical characteristics and geographic location of the home.

(a) (10 pts) Before beginning your data analysis, consider the nature of these variables, how they may interact with each other, and how they may relate to list price. Look at univariate and multivariate summaries, and report any findings that strike you as surprising or important to your inferential goals.

(b) (20 pts) Use a model selection technique to select a model to predict *Price* using the relevant covariates in the dataset. BE CAREFUL. Not all variables in the dataset make sense to use as covariates for the specified inferential goals. Other variables may be so obviously important that you don't need to include them in a variable selection procedure.

(c) (10 pts) Assess deviations from model assumptions. If the assumptions are violated, try to address those concerns and refit the model—or if this isn't possible, explain the ramifications of the violated assumptions in your work.

(d) (5 pts) Write out the final model you select in proper statistical notation, defining parameters carefully. Describe in words the effect of each covariate in the final model.

(e) (5 pts) Use your final model to predict the list price for the following two homes:

Address	Zip Code	SqFt	Beds	Baths	Latitude	Longitude
3001 Calle San Angel NW	87107	2554	3	3	35.124249	-106.676277
10213 Chapala Pl. NE	87111	1475	3	2	35.122654	-106.525421

Point predictions are acceptable, but prediction intervals are preferred.

Problem 2:

The data below come from an experiment on the effectiveness of different adhesive systems.

	Adhesive							
	1	2	3	4	1	2	3	4
With Primer	60	57	19.8	52	73	52	32.0	77
	63	52	19.5	53	79	56	33.0	78
	57	55	19.7	44	76	57	32.0	70
	53	59	21.6	48	69	58	34.0	74
	56	56	21.1	48	78	52	31.0	74
	57	54	19.3	53	74	53	27.3	81
Without Primer	59	51	29.4	49	78	52	37.8	77
	48	44	32.2	59	72	42	36.7	76
	51	42	37.1	55	72	51	35.4	79
	49	54	31.5	54	75	47	40.2	78
	45	47	31.3	49	71	57	40.7	79
	48	56	33.0	58	72	45	42.6	79
	Thickness A				Thickness B			

Here the data recorded in the table are peel strengths recorded under the various adhesive systems. Peel strength is a measure of the amount of force necessary to pull apart the two materials bonded by an adhesive. In these data, pieces of rubber of various thicknesses have been bonded to a surface. In some cases, a primer is applied to the surface before the adhesive; in others, no primer is applied. Four different adhesives are tested in this way.

The dataset contains three covariates:

- *Adhesive* – Which of the four types of adhesive was used in a particular peel strength test.
- *Primer* – Was the adhesive applied to the surface with or without an initial primer coat.
- *Thickness* – Which of two thicknesses of rubber was used in a particular peel strength test.

For each covariate combination, six pieces of rubber are tested, giving six replications for an analysis of variance.

(a) (10 pts) Identify the design for this experiment and give an appropriate model.

(b) (15 pts) Analyze the data. Give an appropriate analysis of variance table and report the simplest linear model that adequately explains these data.

(b) (10 pts) List all the assumptions made by this model, and assess whether there appear to be any deviations from these assumptions. If the assumptions are violated, try to address those concerns and refit the model—or if this isn't possible, explain the ramifications of the violated assumptions in your work.

(d) (5 pts) Consider these data from the perspective of the manufacturer of adhesive #4. You would like to use these data to justify to clients why they should use your adhesive rather than the competing adhesives studied here. Are there any covariates which either must be excluded or must be included in analyses from this perspective? Explain.

(e) (10 pts) If necessary, refit your model to account for your answer to part (d). Then explain the internal recommendations you can make to the marketers of adhesive #4 about the conditions under which the adhesive works best, and how it compares with its competitors.