STAT 425 Exam 2 Study Problem Solutions

Exam problems are generally be shorter than homework problems and may involve short answer conceptual questions, quick calculations and R code interpretation or debugging. It is not a multiple choice exam, although some multiple choice questions are possible.

The sample problems below are to help you test yourself and practice solving. Problems on the exam will generally have fewer parts to them than the ones below. Do not expect the actual exam problems to be exactly like this set in terms of range of coverage or length. Work on these various problems as a way to solidify your understanding.

1. Twenty chicks (baby chickens) were randomly assigned to receive one of two diets, A or B, with 10 in each group. Consider the model

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, 2;\ j = 1, 2, \ldots, 10.$$

Here y_{ij} denotes the 14-day weight gain for the jth chick on Diet i with $i = 1$ for Diet A and $i = 2$ for Diet B. The working model is that the errors are independently normally distributed with mean zero and variance σ^2.

a) Suppose the sample mean responses for the two diet groups are $\bar{y}_A = 101.2$ and $\bar{y}_B = 123.7$. Using the reference category constraint with Diet A as the reference category, calculate the least squares estimates of μ, α_1 and α_2.

$$\hat{\mu} = \bar{y}_A = 101.2$$

$$\hat{\alpha}_1 = 0$$

$$\hat{\alpha}_2 = \bar{y}_B - \bar{y}_A = 123.7 - 101.2 = 22.5$$

b) Calculate the between group sum of squares $FSS = \sum_{i=1}^{2} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$.

$$n_1 = n_2 = 10 \qquad \bar{y}_{1\cdot} = 101.2 \qquad \bar{y}_{2\cdot} = 123.7$$

$$\bar{y}_{\cdot\cdot} = \frac{10 * 101.2 + 10 * 123.7}{20} = \frac{101.2 + 123.7}{2} = 112.45$$

$$FSS = 10 * (101.2 - 112.45)^2 + 10 * (123.7 - 112.45)^2 = 2 * 10 * 11.25^2 = 2531.25$$

c) How many degrees of freedom does FSS have?

$$2 - 1 = 1$$

d) Suppose $\sum_{i=1}^{2}\sum_{j=1}^{10}(y_{ij} - \hat{y}_{ij})^2 = 49.0$. Calculate the value of the F-statistic for testing the null hypothesis $H_0 : \mu_1 = \mu_2 = 0$, where μ_1 is the mean response for Diet A, and μ_2 is the mean response for Diet B.

$$F = \frac{FSS/1}{RSS/(20-2)} = \frac{2531.25}{49/18} = 929.85$$

2. A study was conducted to compare three drug treatments for a certain disease, with drugs labeled A, B and C. For each subject `Pretreatment` is a condition score before treatment. The response `PostTreatment` is the condition score after the treatment regimen. The goal is to determine whether there is a difference between the drugs in improving post-treatment condition, adjusting for the effect of pre-treatment condition.

An analysis of covariance model was fit including the interactions between `Drug` and `Pretreatment`, and the sequential anova table is below.

```
## Analysis of Variance Table
##
## Response: PostTreatment
##                   Df Sum Sq Mean Sq F value    Pr(>F)
## Pretreatment       1 802.94  802.94 48.4726 3.366e-07 ***
## Drug               2  68.55   34.28  2.0692    0.1482
## Pretreatment:Drug  2  19.64    9.82  0.5930    0.5606
## Residuals         24 397.56   16.56
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a) What is the overall sample size, n?

$$5 + 24 + 1 = 30$$

b) What hypothesis is being tested by the following row in the table, and what do you conclude from the result?

```
Pretreatment:Drug 2 19.64 9.82 0.5930 0.5606
```

Equivalent statements:

H_0 : No interaction between Pretreatment and Drug effects

H_0 : The additive model is adequate: `PostTreatment` $\sim$ `Pretreatment + Drug`

c) Based on the sequential anova results what is the best model for these data:

`PostTreatment` $\sim$ `1`

`PostTreatment` $\sim$ `Pretreatment`

`PostTreatment` $\sim$ `Pretreatment + Drug`

`PostTreatment` $\sim$ `Pretreatment + Drug + Pretreatment:Drug`

Explain why.

`PostTreatment` $\sim$ `Pretreatment`. Reason: stepping backward from the full interaction model, the interaction terms are not significant with the two main effects in the model, and, dropping the interaction, the main effect for `Drug` is not significant with `Pretreatment` in the model.

d) Write out the model formula (in R syntax) for the model that corresponds to three parallel regression lines for PostTreatment versus Pretreatment for the three Drug groups.

`PostTreatment ~ Pretreatment + Drug`

e) Consider the following notation to express the variables in the data mathematically for the ith subject:

y_i is the PostTreatment score,

x_i is the Pretreatment score,

$z_{i1} = 1$ if Drug A and $z_{i1} = 0$ if not Drug A,

$z_{i2} = 1$ if Drug B and $z_{i2} = 0$ if not Drug B,

$z_{i3} = 1$ if Drug C and $z_{i3} = 0$ if not Drug C,

and e_i is the error term. Using this notation, write out a valid, full rank mathematical form of the model corresponding to the R formula:

`PostTreatment ~ Pretreatment + Drug + Pretreatment:Drug`

Use expressions like β_0, β_1 etc. for the coefficients of the model.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{i2} + \beta_3 z_{i3} + \beta_4 x_i z_{i2} + \beta_5 x_i z_{i3} + e_i, \quad i = 1, 2, \ldots, 30$$

Note that we only use two of the three indicator variables to distinguish the three Drug categories. We can tell it's Drug A if $z_{i2} = z_{i3} = 0$. So we make Drug A the reference value, and β_2 and β_3 are incremental effects of Drugs B and C, respectively, versus Drug A. The interactions are coded as products of the Pretreatment and Drug variables.

3. A cubic polynomial was fit using the crossx variable as the response and energy as the predictor. The sequential ANOVA table for this fitted model is as follows:

```
## Analysis of Variance Table
##
## Response: crossx
##             Df  Sum Sq Mean Sq   F value    Pr(>F)    
## energy       1 272.216 272.216 1265.6028 3.289e-08 ***
## I(energy^2)  1  10.980  10.980   51.0492 0.0003789 ***
## I(energy^3)  1   0.622   0.622    2.8933 0.1398498    
## Residuals    6   1.291   0.215                        
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a) Based on the numerical results above, select one of the following options to describe the most appropriate model for this data set, and explain your choice:

1. A linear trend or simple linear regression of `crossx` on `energy`.
2. A quadratic polynomial model.
3. A cubic polynomial model.
4. We do not have enough information to select among the options above.

   2. A quadratic polynomial model. Reason: Starting from the bottom of the sequential anova table, the cubic term is not statistically significant, so we would drop it. The quadratic term `I(energy`^{2}`)` is highly significant so we keep it and stop.

b) In the sequential ANOVA table above, the F test corresponding to the quadratic term `I(energy`^{2}`)` is calculated as the ratio of two numbers

1. Numerator: A= (use a number with two digits after the decimal)
2. Denominator: B= (use a number with two digits after the decimal)

Give the numerator A, denominator B, and degrees of freedom for this F test.

$$A = 10.98 \text{ (I(energy}^2\text{) Mean Sq)}, \quad B = 0.215 \text{ (Residuals Mean Sq)}$$

$$df = 1 \text{ and } 6 \text{ (numerator and denominator)}$$

c) Consider the following notation to express the variables mathematically for the ith observation: y_i is the value for `crossx`, x_i is the value for `energy`, and e_i is the error. Using this notation, write out a valid mathematical form of the model corresponding to the cubic polynomial model. For the coefficients, use expressions like β_0, β_1, etc.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + e_i,\ i = 1, 2, \ldots, 10$$

4. Each of the following R function calls create a set of basis functions. For each, give the degrees of freedom and the total number of knots. Note - problem should have specified that the lm function will include the intercept if it's not already in the basis functions.

a) B-spline:

```
bs(year, df=6, intercept=TRUE)
```

df: 6

knots: 6-4=2

b) B-spline:

```
bs(year, df=8, intercept=FALSE)
```

df: 9 assuming the intercept will be included in the model, e.g. by `lm`

knots: 9-4 = 5

c) Natural Cubic Spline:

```
ns(year, df=8, intercept=TRUE)
```

df: 8

knots: 8-2=6

d) Natural Cubic Spline:

```
ns(year, df=10, intercept=FALSE)
```

df: 10+1=11

knots: 11-2=9

5. The following output summarizes the options for the first step in a backwards stepwise selection process applied to a linear model for data relating mean life expectancy for U.S. states to other demographics. The first column is the list of candidate variables for deletion. The RSS column shows the residual sum of squares for the model without the indicated variable, and the AIC column shows the AIC for that model.

```
## Single term deletions
##
## Model:
## Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
##     Frost + Area
##            Df Sum of Sq    RSS     AIC
## <none>                  23.297 -22.185
## Population  1    1.7472 25.044 -20.569
## Income      1    0.0044 23.302 -24.175
## Illiteracy  1    0.0047 23.302 -24.174
## Murder      1   23.1411 46.438  10.305
## HS.Grad     1    2.4413 25.738 -19.202
## Frost       1    1.8466 25.144 -20.371
## Area        1    0.0011 23.298 -24.182
```

a) Which, if any, variable would be removed if we use AIC as the selection criterion for backward stepwise regression? Explain why.

`Area`: dropping this variable gives the lowest AIC of -24.182

b) What is the AIC value for the model that includes all variables?

-22.185

c) The full model we started with has smaller RSS than any of the candidate models obtained by dropping one variable. Does this mean that the full model is actually the best option? Explain why or why not.

No. RSS only measures fit, not complexity. We can driving down the RSS by adding more variables, but this leads to over-fitting the particular data set. The resulting model would not be a good predictive model for new data. AIC is more reliable because it prevents over-fitting by penalizing complexity.

6. Several questions about regularized regression.

a) When we use principal components regression, it is always best to use all principal components of the design matrix **X** for regression. True or False? Explain.

> False. This would be equivalent to using all of the variables in **X**. We will obtain better predictions by reducing to a smaller set of principal components that account for a large percentage of the variation in **X**.

b) Consider the following output after computing the principal components of predictor variables:

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6    PC7
## Standard deviation     1.7548 1.2739 1.0025 0.74634 0.58222 0.50886 0.3713
## Proportion of Variance 0.4399 0.2318 0.1436 0.07958 0.04843 0.03699 0.0197
## Cumulative Proportion  0.4399 0.6717 0.8153 0.89488 0.94331 0.98030 1.0000
```

According to this output, how many principal components are needed to account for at least 90% of the variation in the predictor variables? Give the actual percentage of variation accounted for by this choice.

> 5 principal components. From the "Cumulative Proportion" row the first 5 principal components account for **94.3%** of the variation, whereas the first 4 account for slightly less than 90%.

c) Lasso regression minimizes the residual sum of squares of a regression model subject to the constraint that $\sum_{j=1}^{p} |\beta_j| \leq t$. Which of the following methods could be used to select the value for t:

1. Maximum likelihood
2. Least squares
3. Weighted least squares
4. Cross-validation
5. None of the above

> 4. Cross-validation. This provides an unbiased way to estimate which constrained model gives the best out-of-sample predictions.

7. The following results are from fitting a one-way anova model to data relating blood coagulation times to diet.

```
##
## Call:
## lm(formula = coag ~ diet, data = coagulation)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5.00  -1.25   0.00   1.25   5.00
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.100e+01  1.183e+00  51.554  < 2e-16 ***
## dietB       5.000e+00  1.528e+00   3.273 0.003803 **
## dietC       7.000e+00  1.528e+00   4.583 0.000181 ***
## dietD       2.991e-15  1.449e+00   0.000 1.000000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.366 on 20 degrees of freedom
## Multiple R-squared:  0.6706, Adjusted R-squared:  0.6212
## F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

a) Based on these results, compute the estimated difference between the mean coagulation time for Diet B and the mean for Diet A. Also give the standard error for this difference if possible.

We can see from the results that Drug A is the reference group, so this mean difference is estimated by the `dietB` coefficient. Estimate = 5.00, Standard Error = 1.53

b) Below is the analysis of variance table for the model above.

```
## Analysis of Variance Table
##
## Response: coag
##           Df Sum Sq Mean Sq F value    Pr(>F)
## diet       3    228    76.0  13.571 4.658e-05
## Residuals 20    112     5.6
```

State the null hypotheses that the F value is testing. Does the test reject the null hypothesis at level 0.05?

$H_0 : \mu_A = \mu_B = \mu_C = \mu_D$ (all treatment group means are equal)

p $<$ .0001 so the test rejects the null hypothesis at level 0.05.

c) Based on the results below, determine which pairs of diet group means are significantly different from each other, controlling the family wise error rate at $\alpha = 0.05$.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = g)
##
## $diet
##     diff         lwr       upr     p adj
## B-A    5   0.7245544  9.275446 0.0183283
## C-A    7   2.7245544 11.275446 0.0009577
## D-A    0  -4.0560438  4.056044 1.0000000
## C-B    2  -1.8240748  5.824075 0.4766005
## D-B   -5  -8.5770944 -1.422906 0.0044114
## D-C   -7 -10.5770944 -3.422906 0.0001268
```

The following pairs of mean differences have adjusted p-values < 0.05 and are therefore statistically significant:

$$B - A\,(> 0), \quad C - A\,(> 0), \quad D - B\,(< 0), \quad D - C\,(< 0)$$

Another way to think about this result is that the means are grouped as follows:

$$\{A, D\} < \{B, C\}$$

with significant differences between the two bracketed groups, and no significant differences within the bracketed groups.

8. Consider the following linear model assumptions.

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}, \quad E(\mathbf{e}) = \mathbf{0}, \quad Cov(\mathbf{e}) = \sigma^2\mathbf{V},$$

where $\mathbf{V}$ is a diagonal matrix of the form

$$\mathbf{V} = \begin{pmatrix} v_1 & 0 & 0 & \cdots & 0 \\ 0 & v_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & v_n \end{pmatrix}$$

and the diagonal elements are all positive.

a) Let $\mathbf{W}$ be the diagonal matrix

$$\mathbf{W} = \begin{pmatrix} \frac{1}{\sqrt{v_1}} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{v_2}} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & \frac{1}{\sqrt{v_n}} \end{pmatrix}$$

Show that the transformed response vector $\mathbf{Z} = \mathbf{WY}$ follows a modified linear model in which the error vector has mean $\mathbf{0}$ and covariance $\sigma^2\mathbf{I}_n$.

Using the model equation we have

$$\mathbf{Z} = \mathbf{WY} = \mathbf{WX}\beta + \mathbf{We} = \mathbf{X}^*\beta + \mathbf{e}^*,$$

where $\mathbf{X}^* = \mathbf{WX}$ and $\mathbf{e}^* = \mathbf{We}$. Furthermore, $E(\mathbf{e}^*) = WE(\mathbf{e}) = 0$ and $Cov(\mathbf{e}^*) = \sigma^2\mathbf{WVW^T} = \sigma^2\mathbf{I}$.

b) The weighted sum of squared residuals for the original model has the form

$$RSS_w(\beta) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)^2/v_i = (\mathbf{Y} - \mathbf{X^T}\beta)\mathbf{WW}(\mathbf{Y} - \mathbf{X^T}\beta)$$

Show that minimizing $RSS_w(\beta)$ as a function of β is the same as miminizing the ordinary (unweighted) residual sum of squares as a function of β for the modified linear model for $\mathbf{Z}$.

First note that $\mathbf{W^T} = \mathbf{W}$ and $\mathbf{WW} = \mathbf{V}^{-1}$. The OLS residual sum of squares for the transformed model in a) is given by

$$\begin{aligned} RSS^*(\beta) &= (\mathbf{Z} - \mathbf{X}^*\beta)^T(\mathbf{Z} - \mathbf{X}^*\beta) \\ &= (\mathbf{WY} - \mathbf{WX}\beta)^T(\mathbf{WY} - \mathbf{WX}\beta) \\ &= (\mathbf{W}(\mathbf{Y} - \mathbf{X}\beta))^T(\mathbf{W}(\mathbf{Y} - \mathbf{X}\beta)) \\ &= (\mathbf{Y} - \mathbf{X}\beta)^T\mathbf{W}^T\mathbf{W}(\mathbf{Y} - \mathbf{X}\beta) \\ &= (\mathbf{Y} - \mathbf{X}\beta)^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\beta) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)^2/v_i \end{aligned}$$

Therefore, minimizing $RSS_w(\beta)$ is the same as minimizing $RSS^*(\beta)$.