Statistics 203
Introduction to Regression Models and ANOVA
Practice Exam

Prof. J. Taylor

You may use your 4 single-sided pages of notes

This exam is 7 pages long. There are 4 questions, first 3 worth 10 points, the fourth worth 15 points.

Q. 1) A model is fit to predict `MPG` (miles per gallon) of several makes of cars, based on `WT` (weight), `SP` (speed); `VOL` (cab volume) and `HP` (horse power). Here is the output of R's `summary` for that model.

```
> lm1 = lm(MPG ~ WT+VOL+HP+SP)
> summary(lm1)

Call:
lm(formula = MPG ~ WT + VOL + HP + SP)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0108 -2.7731  0.2733  1.8362 11.9854

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.43775   23.53161   8.178 4.62e-12 ***
WT           -1.85980    0.21336  -8.717 4.22e-13 ***
VOL          -0.01565    0.02283  -0.685    0.495
HP            0.39221    0.08141   4.818 7.13e-06 ***
SP           -1.29482    0.24477  -5.290 1.11e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.653 on 77 degrees of freedom
Multiple R-squared: 0.8733,Adjusted R-squared: 0.8667
F-statistic: 132.7 on 4 and 77 DF,  p-value: < 2.2e-16
```

(a) Which entry above is $\widehat{\sigma}$, i.e. the usual estimate of the parameter σ in this model?

(b) What is the error sum of squares for this model, i.e. SSE? (No need to compute this with a calculator, just give an expression using entries from the table).

(c) If you knew the SSE, how would you compute the total sum of squares of MPG, i.e. SST, from the above output?

(d) Suppose you just wanted to test whether VOL is unrelated to MPG allowing for all the other effects in the model. Test this hypothesis with an F test. What is the p-value of this test?

(e) Finally, the research team has decided that they want to test the hypothesis that the coefficients for HP and VOL are zero. That is, that HP and VOL are unrelated to MPG when allowing or controlling for the other variables in the model. To do so, they fit another model. Here is R's `summary` of that model.

```
> lm2 = lm(MPG ~ WT+SP)
> summary(lm2)

Call:
lm(formula = MPG ~ WT + SP)

Residuals:
     Min       1Q   Median       3Q      Max
-10.5160  -2.5085  -0.8544   0.9377  16.6276

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 75.64938    3.89181  19.438   <2e-16 ***
WT          -0.99738    0.07774 -12.830   <2e-16 ***
SP          -0.09816    0.04508  -2.177   0.0325 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.184 on 79 degrees of freedom
Multiple R-squared: 0.8294,Adjusted R-squared: 0.8251
F-statistic: 192.1 on 2 and 79 DF,  p-value: < 2.2e-16
```

Describe how you could test this hypothesis using the above output.

	Df	SSE
Use	2	33.24
Size	3	12.06
Size:Use	6	4.25
Residuals	108	154.62

Table 1: ANOVA output for gasoline consumption analysis

Q. 2) In order to determine some of the factors that effect gasoline consumption, the fuel average miles per gallon for 120 families and their cars were tracked over several months. The cars were categorized in to 4 categories: compact, sedan, minivan, SUV; and the prevalent use was broken down into 3 categories: commuting to work, weekend, driving to school. The sums of squares of a two-way analysis of variance (ANOVA) model are presented above.

(a) What assumptions does the two-way ANOVA model make? Be as precise as possible.

(b) Compute the F-statistic used to test for the main effects of both Use and Size.

(c) Based on the definition of the F distribution, if the denominator degrees of freedom of an F-statistic is large, what might you use to approximate the expected value of the F-statistic under the appropriate null hypothesis? Using that guideline, does there appear to be evidence for main effects of Use and Size?

(d) Compute the F-statistic to test for an interaction between Use and Size. Using the above guideline, does there appear to be evidence for an interaction effect?

Q. 3) A liver specialist has come to you data relating the number of alcoholic beverages per week for each subject to the chances that they have develop liver disease (within some long follow-up period). The study controlled for the effects of `Age` as well as overall fitness level `Fitness` with a high value indicating a very fit subject. The results of a logistic regression used in the study were:

```
Call:
glm(formula = Y ~ Age + Drinks + Fitness, family = binomial(link='logit'))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.21798  -0.26941  -0.16629  -0.08747   2.44946

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.8436     5.1508   0.552  0.58090
Age          -0.0993     0.1028  -0.966  0.33418
Drinks        0.1601     0.1208   1.325  0.18503
Fitness      -0.7584     0.2411  -3.146  0.00166
```

(a) What is the estimated probability that a 50-year old who drinks 5 drinks a week of fitness level 3 will develop liver disease? Redo this calculation for a 50-year old of the same fitness level who does not drink any alcoholic beverages? (If you don't have a calculator, just don't simplify the expression).

(b) Using an odds ratio, approximate how many more times likely is the adult who drinks 5 drinks likely to develop liver disease than the person who drinks none?

(c) The liver specialist guesses that there might be some interaction between fitness level and the number of drinks per week on the chances of liver disease. Describe how you might fit a model that allows for this interaction and use this model to test the hypothesis (in approximately correct `R` code).

(d) Suppose the specialist had decided to discretize `Fitness` into 3 levels, say (L,M,H). How would the rows of the `summary` coefficients table above be different? What about the rows in the model for part (c)? (Don't worry if you do not guess exactly what 'R' would do, the general idea is sufficient.)

Q. 4) Consider a regression model

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \delta_{n\times 1}.$$

Unlike the usual assumptions, though, in this model it is assumed that nature has conspired against us in that $E(X^T\delta) \neq 0$ though β is the true parameter of interest. Suppose we also have access to other variables $Z_{n\times q}$ for which $E(Z^T\delta) = 0$. (This is the so-called *instrumental variables* model.)

The instrumental variables estimate for β is

$$\hat{\beta}_{IV} = (X^T Z(Z^T Z)^{-1} Z^T X)^{-1}(X^T Z(Z^T Z)^{-1} Z^T Y). \tag{1}$$

The estimator $\hat{\beta}_{IV}$ can be motivated from the equation

$$Z^T Y = (Z^T X)\beta + Z^T\delta \tag{2}$$

and is the *generalized least-squares estimator* assuming that $\mathrm{Cov}(Z^T\delta) = \sigma^2 Z^T Z$.

(a) What do we need to assume about p, q above so that $\hat{\beta}_{IV}$ is well-defined?

(b) Pretending that $X^T Z$ and $Z^T Z$ are constant, show that $E(\hat{\beta}_{IV}) = \beta$.

(c) Assume that the cases (X_i, Z_i, δ_i) for $1 \leq i \leq n$ are IID from some distribution F with δ_i independent of Z_i with $E_F(\delta_1) = 0$. What does the law of averages say about $X^T Z$ and $Z^T Z$ as $n \to \infty$?

(d) What does the law of averages say about $Z^T\delta$ as $n \to \infty$? What is the covariance of $Z^T\delta$? (Recall that Z_i is independent of each δ_i.)

(e) Use parts (c) and (d) above to conclude that $\hat{\beta}_{IV}$ is well-approximated in distribution by

$$N(\beta, \mathrm{Var}_F(\delta)(X^T Z(Z^T Z)^{-1} Z^T X)^{-1}).$$

(In practice, the quantity $\mathrm{Var}_F(\delta)$ can be estimated by $\|Y - X\hat{\beta}_{IV}\|^2/(n-p)$.

(f) Another estimator we might use upon inspection of (2) is the OLS estimator in this equation:

$$\tilde{\beta} = (X^T Z Z^T X)^{-1} X^T Z Z^T Y.$$

Show that $E(\tilde{\beta}) = \beta$. How might you approximate its distribution?