Statistics 191
Introduction to Regression Analysis and Applied Statistics
Practice Exam # 2

Prof. J. Taylor

You may use your 4 single-sided pages of notes
This exam is 8 pages long. There are 4 questions, each worth 10 points.

I understand and accept the Stanford University Honor Code.

Name: _____

Signature: _____

1	
2	
3	
4	
Total	

(Intercept)	Treat2	Treat3
1	0	0
1	0	0
⋮	⋮	⋮
1	1	0
1	1	0
1	1	0
⋮	⋮	⋮
1	0	1
1	0	1
1	0	1
1	0	1

Table 1: Part of design matrix for R output.

Q. 1) A coffee grower wants to determine if using different agricultural techniques on his coffee beans will significantly affect the amount of caffeine per cup of coffee produced from his beans. He tries three different treatments and measures the caffeine content under each technique, resulting in the R output below.

```
Call:
lm(formula = Caffeine ~ Treat)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9430 -1.3928  0.2376  1.5879  3.6151

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 109.8619     0.9802  -0.879 0.396484
Treat2        3.4601     1.3862   2.496 0.028120 *
Treat3        6.7752     1.3862   4.887 0.000374 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.192 on 12 degrees of freedom
Multiple R-squared: 0.6657,Adjusted R-squared: 0.6099
F-statistic: 11.95 on 2 and 12 DF,  p-value: 0.001397
```

You are also given some of the design matrix R used in fitting this model in Table 1.

(a) The grower repeated his experiment on a number of different fields. If the grower had the same number of fields per treatment (i.e. his

design was balanced), how many fields per treatment did he use?

There are 12 residual degrees of freedom and we estimated 3 parameters. It was 5 fields per treatment.

(b) What was the mean amount of caffeine in fields treated with Treatment 1? Treatment 2? Treatment 3?

1: 109.8, 2: 109.8+3.4, 3: 109.8+6.77

(c) What can you conclude about the average amount of caffeine change when using different treatments? Give your answer in terms of a statistical test. What is the null hypothesis? The alternative? The p-value?

This is a one way ANOVA model. We can test the null hypothesis that the amount of caffeine is the same under all three treatments with the overall F test. The F statistic is 11.95 on 2 and 12 degrees of freedom, p-value less than 5% so at level 5% we can reject the null hypothesis that the caffeine is the same.

(d) If you had to tell a friend how to interpret the p-value of this test in the output above, what would you say? You can use the grower's experiment to describe its interpretation if it helps.

This p-value can be used to test the null hypothesis of no difference. It is a measure of the evidence against the null hypothesis. If there really was no difference between the treatments, and we ran this experiment many times, then only about 5% of the time would our p-value be less than 5%.

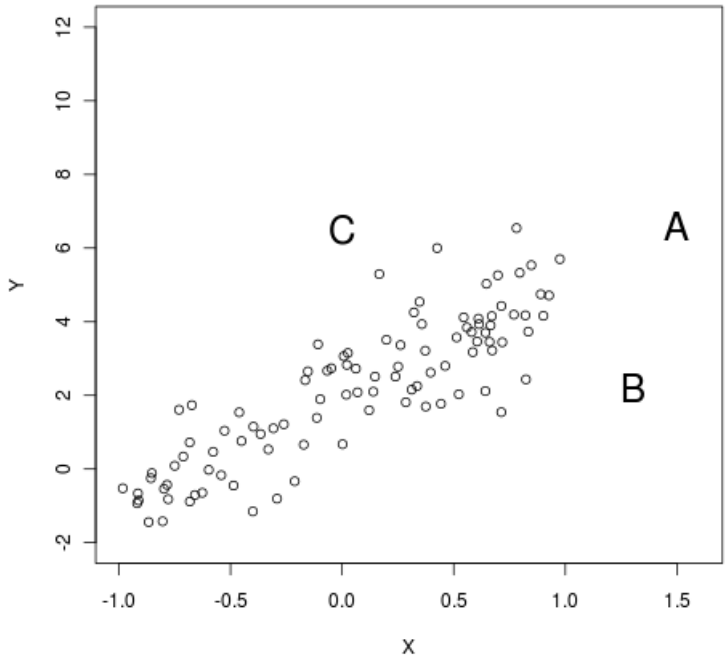

Figure 1: Figure for Q. 3)

Q. 2) Figure 1 shows the data from a regression study with only one predictor. In each question, the correct answer may be more than one of A, B or C.

(a) Give a definition of *leverage*. Which of the labelled points (A, B or C) would you think have a high leverage value? Explain.
A, B

(b) Which of the labelled points (A, B or C) would you think would be labelled as an outlier by an outlier detection test? Explain. B or C

(c) What does Cook's Distance try to measure for a multiple linear regression model? Which of the labelled points (A, B or C) would you think would have a large Cook's distance? Explain.

B or C

Q. 3) Figure 2 shows the diagnostic plots from a model fit predict `MPG` (miles per gallon) of several makes of cars, based on `WT` (weight), `SP` (speed); `VOL` (cab volume) and `HP` (horse power).

(a) Briefly describe the assumptions used in fitting this model. Do any of the assumptions seem to be violated based on Figure 3? If any seem to be violated, suggest some possible fixes.

Our model is

$$MPG_i = \beta_0 + \beta_{WT} WT_i + \beta_{SP} SP_i + \beta_{VOL} VOL_i + \beta_{HP} HP_i + \epsilon_i$$

where the ϵ_i's are independent errors with normal distributions $N(0, \sigma^2)$.

From the diagnostic plots, we see that perhaps the errors are non-normal from the `qqplot`. The variance of the error may not be quite constant as well. Fixing the errors to be normal is difficult, particularly if the regression function is (close to) correct. It is possible that non-constant variance is manifesting itself in non-normal looking errors.

It might be possible to model the variance of the errors to adjust their variance.

(b) As part of the output, the researchers computed both a 95 % confidence interval and a 95 % prediction interval for the `MPG` of a car with `WT=30, HP=120, SP=110, VOL=100`, only they can't remember which of the following two intervals was which.

```
          fit      lwr      upr
[1,] 39.71459 37.56043 41.86874
>
          fit      lwr      upr
[1,] 39.71459 32.12874 47.30044
>
```

Which interval is the confidence interval? Explain.

The confidence interval will be shorter because the prediction interval needs to cover a new draw from the error distribution.

(c) Your friend now says:

> *So this means that if you repeat this experiment 100 times, roughly 95 times the true mean weight of a car with* `WT=30, HP=120, SP=110, VOL=100` *will be between the limits of the confidence interval you chose in (b) (either [32.1,47.3] or [37.5,41.9]).*

Do you agree? Why or why not?

We disagree. If we collected similar data many times and formed many different confidence intervals, about 95% of them would cover the true mean for such a car. The true mean is a fixed number. It is contained in the interval[47.5,41.9] or not – there are no chances to compute here.

(d) As part of this project, the researcher also wants to find a model that fits the model possibly better than this current model. The researcher is willing to consider some potential interactions in the model. Describe some possible approaches, being as explicit as reasonably possible (use `R` pseudo-commands if you like).

We might try forward stepwise with 'step'. Or perhaps LASSO including all possible interactions. Fairly open ended.

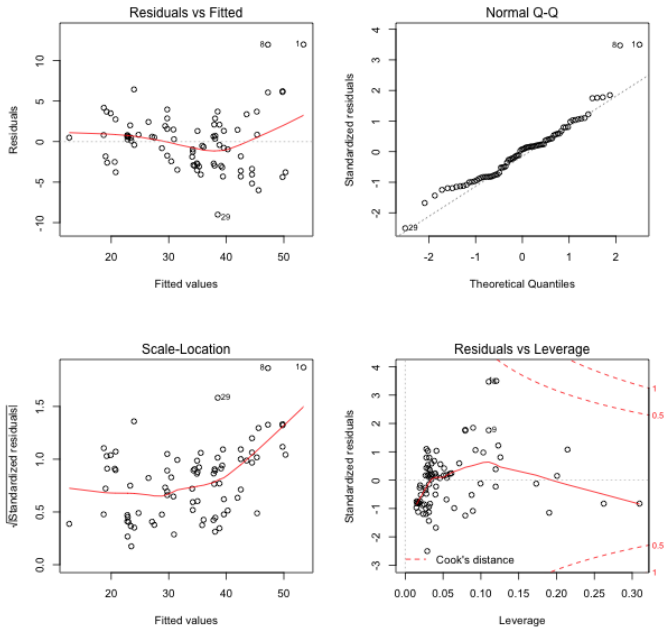

Figure 2: Diagnostic plots for regression of MPG (miles per gallon) of several makes of cars, based on `WT` (weight), `SP` (speed); `VOL` (cab volume) and `HP` (horse power).

Q. 4) 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? (If you don't have a calculator, just don't simplify the expression).

(b) 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).

(c) 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?

(d) Construct approximate Bonferroni-adjusted 90% confidence intervals for the coefficients of `Age, Drinks, Fitness`. (See below for some potentially useful R output)

Some potentially useful R output for Q. 4):

```
> qnorm(1-0.1)
[1] 1.281552
> qnorm(1-0.1/2)
[1] 1.644854
> qnorm(1-0.1/3)
[1] 1.833915
> qnorm(1-0.1/6)
[1] 2.128045
>
```