Statistics 191
Introduction to Regression Analysis and Applied Statistics
Practice Exam

Prof. J. Taylor

You may use your 4 single-sided pages of notes

This exam is 14 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	

Q. 1) In order to study the relevance of different fields of study in the job market, the Stanford alumni association followed Stanford graduates in the first few years after graduation. Students were asked their starting salaries, as well as which sector (i.e. finance, technology, education) they were employed in. Within the finance field there were (among others) (20 Math & Computational Science (MCS) & 20 English graduates).

(a) The first question the researchers addressed was whether an MCS degree was worth more than a English degree in finance. After entering the data in R, they found the following:

```
> t.test(mcs, english, var.equal=T, alternative='greater')

        Two Sample t-test

data:  mcs and english
t = 3.718, df = 38, p-value = 0.0003228
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 5509.375      Inf
sample estimates:
mean of x mean of y
 69457.52  59377.05
```

Explain the results in the output above and what conclusions the researchers can draw.

The above results are from a two-sample T-test testing whether the average salary is the same within MCS graduates as English graduates. The researchers conducted a one-sided test in which the null hypothesis was

$$H_0 : \mu_{MCS} \leq \mu_{English}$$

and the alternative was

$$H_a : \mu_{MCS} > \mu_{English}.$$

The t statistic is computed as

$$T = \frac{\hat{\mu}_{MCS} - \hat{\mu}_{English}}{SE(\hat{\mu}_{MCS} - \hat{\mu}_{English})}$$

and was observed to be 3.718, much larger than 0. This is strong evidence against H_0, with a p-values of about $3 * 10^{-4}$. We reject H_0 and conclude H_a is true.

(b) A friend of your reads the results of this study and says:

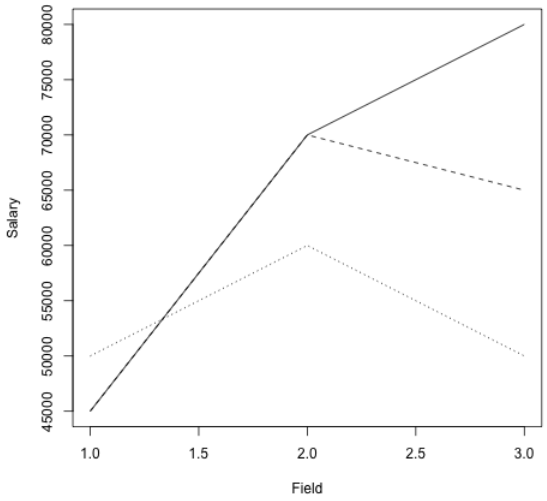
> *Wow! If I choose MCS as my major, the probability I'd earn more than a English grad if we both start in finance is about* $1 - 3 * 10^{-4}$*!*

Do you agree with your friend's statement? Explain.

We should not agree. The probability above is the probability a T statistic with the appropriate degrees of freedom would be as large as the one we observed. This is not the probability the null hypothesis is true, which is the error our friend has made.

(c) How would you obtain the same results in (a) using the `lm` command in `R`? Approximately correct syntax is OK here.

```
group = factor(c(rep("MCS",20), rep("English",20)))
Y = c(mcs, english)
summary(lm(Y ~ group))
```



Q. 2) The study begun in Q. 1) was continued, expanding to several fields: education, finance and technology; as well as an additional degree: electrical engineering. The means in each group were (after rounding)

	Education	Finance	Technology
EE	45000	70000	80000
MCS	45000	70000	65000
English	50000	60000	50000

(a) Sketch the *interaction plot* for this model, depicting the data in the above table. From this plot, does the type of degree affect your starting salary? What about the field you begin work in? Are there any interactions? Explain.

We have used 1=Education, 2=Finance, 3=Technology in the x-axis above. As the graphs are not flat, we conclude there is evidence for a main effect of field. As the graphs do not lie one on top of each other, we concluded there is evidence for a main effect of degree earned. As the graphs are not parallel, we see there is evidence of interactions between the field one works in and the degree earned.

(b) The output below results from fitting this model. What kind of a regression model is it?

```
> anova(lm(Salary~Degree*Field))
Analysis of Variance Table

Response: Salary
              Df     Sum Sq   Mean Sq  F value
Degree         2 6.2969e+09 ????????? ???????
Field          2 1.7624e+10 ????????? ???????
Degree:Field   4 6.7457e+09 ????????? ???????
Residuals    171 1.4860e+10 ?????????
---
```

Make a table that includes all values overwritten with ?'s above. (Don't worry if you don't have a calculator, fractions are fine.)

This is a 2-way ANOVA model. Here is the completed table:

```
> anova(lm(Salary~Degree*Field))
Analysis of Variance Table

Response: Salary
              Df     Sum Sq   Mean Sq             F value
Degree         2 6.2969e+09 6.2969e+09/2 (6.2969e+09/2) / (1.4860e+10/171) = 36.2
Field          2 1.7624e+10 1.7624e+10/2 (1.7624e+10/2) / (1.4860e+10/171) = 101.4
Degree:Field   4 6.7457e+09 6.7457e+09/4 (6.7457e+09/4) / (1.4860e+10/171) = 19.4
Residuals    171 1.4860e+10 1.4860e+10/171
---
```

(c) How would you compute p-values for each entry of the `F value` column if you had `R`? Without explicitly computing p-values, do the F statistics support your conclusions in (a)? State the hypothesis that each `F value` is testing. YOU CAN USE THE FACT THAT, WITH MANY DEGREES OF FREEDOM IN THE DENOMINATOR, AN F STATISTIC HAS EXPECTED VALUE 1 IF THE NULL HYPOTHESIS IS TRUE.

The p-values would be

```
1 - pf(36.2, 2, 171)
1 - pf(101.4, 2, 171)
1 - pf(19.4, 4, 171)
```

As all of these F-statistics are much larger than 1, which is roughly the average of an F statistic with 171 degrees of freedom, we conclude that we would conclude there are main effects as well as an interaction effect.

Q. 3) The incidence of landslides in Northern California increases with the amount of rain that falls in any given year. Suppose we are given data of the following form, collected for each month over several years.

`Landslides`: The number of landslide events reported in the Santa Cruz mountains in that specific month.

`Rainfall`: The average rainfall in the Santa Cruz mountains in that month measured in inches.

`RainfallP`: The average rainfall in the Santa Cruz mountains in the previous month measured in inches.

(a) Consider the following `R` output

```
MA = lm(Landslides ~ Rainfall + RainfallP)
MA
##
## Call:
## lm(formula = Landslides ~ Rainfall + RainfallP)
##
## Coefficients:
## (Intercept)     Rainfall    RainfallP
##      0.6699       0.1065       0.0274

MB = glm(Landslides ~ Rainfall + RainfallP, family = poisson())
MB
##
## Call:  glm(formula = Landslides ~ Rainfall + RainfallP, family = poisson())
##
## Coefficients:
## (Intercept)     Rainfall    RainfallP
##      0.1807       0.0414       0.0103
##
## Degrees of Freedom: 119 Total (i.e. Null);  117 Residual
## Null Deviance:     136
## Residual Deviance: 117  AIC: 436
```

Which model would you think to be the most natural model to model the count of the number of landslide events in the Santa Cruz mountains? Explain.

As we are counting earthquakes, which are events a Poisson model seems more appropriate. In the Poisson model, the default link is the log link meaning things act multiplicatively. This is model `MB`.

(b) In whichever model you chose, what is the effect of receiving an additional 10 inches of rain in a given month?

We chose the Poisson model (`MB`) which, when fit, estimates that for every 10 inches of rain in a given month the expected number of earthquakes is increased multiplicatively by a factor of $e^{10*0.0414} = e^{0.414}$. If we had chosen model (`MA`), we would estimate that the number of earthquakes would increase linearly by an amount of $10 * 0.1065 = 1.065$.

(c) How would you use `anova` to test whether the variable `RainfallP` is associated to `Landslides`. What is the null hypothesis in your test? The alternative?

Continuing with model (`MB`), we could test this with

```
anova(glm(Landslides ~ Rainfall, family=poisson()), MB)
```

(d) Suppose you hypothesize that there is a *supersaturation* effect. That is, when rainfall exceeds 15 inches in a given month, the relationship between `Rainfall` and `Landslides`. You decide to form a variable

```
HeavyRainfall = (Rainfall > 15)
```

Write the formula for a model that allows for a different slope and intercept in months of heavy rainfall. How many degrees of freedom does this model use to estimate the mean? (ASSUME THAT YOU HAVE RETAINED THE VARIABLE `RainfallP` FROM ABOVE.)

The model

```
glm(Landslides ~ HeavyRainfall * Rainfall + RainfallP)
```

allows for different slopes and intercepts for the "lines" in months of heavy vs. light rainfall. This model has 5 parameters: `(Intercept)`, `Rainfall`, `HeavyRainfall`, `HeavyRainfall:Rainfall`, `RainfallP` so we would say it has used 5 degrees of freedom to estimate the mean.

Gbar	N	T
30.4	7	80
32.1	4	90
36.2	9	100
33.5	3	110
38.4	11	120

Table 1: Data on averaged growth rate at different temperatures T. The column *Gbar* is the average of N different experiments at a fixed temperature T.

Q. 4) A lab scientist is interested in the effect of temperature, T on the growth rate of a certain population of bacteria. For each of the temperatures in the Table 1 below, the scientist performs a number of experiments, which can be found in the column N, and records the average growth rate (averaged across all experiments of that temperature) in the column `Gbar` of Table 1. That is, 30.4 is the average of growth rates, G, from 7 different experiments each conducted at a temperature of 80. The scientist is interested in fitting a linear regression model for the growth rate as a function of T.

Assume that a simple linear model of the form

$$G = \beta_0 + \beta_1 T + \epsilon \tag{1}$$

is appropriate if the actual growth rates were observed (rather than `Gbar`, their averages over the N experiments). That is, the variance of ϵ is constant across different values of T and is distributed as $N(0, \sigma^2)$ independently for each experiment.

(a) The quantities `Gbar` are averages over several experiments. What is the relationship between the variances of the `Gbar` entries in Table 1 and σ^2 in (1)? Is their variance constant if (1) is correct?

The relationship is

$$\text{Var}(\texttt{Gbar}_i) = \frac{1}{N_i}\text{Var}(\epsilon).$$

The variance is therefore not constant over the cases.

(b) Your are given a choice between the following 3 outputs from R to find a confidence interval for the coefficient β_1 in the model (1).

```
> # MODEL A
> confint(lm(Gbar~T, growth))
                  2.5 %     97.5 %
(Intercept) -2.45005910 35.8900591
T           -0.01581187  0.3638119
```

```
> # MODEL B
> confint(lm(Gbar~T, growth, weights=growth$N))
                2.5 %     97.5 %
(Intercept) 0.13407264 31.3517067
T           0.03736585  0.3399493


> # MODEL C
> confint(lm(Gbar~T, growth, weights=1/growth$N))
                 2.5 %     97.5 %
(Intercept) -1.07474463 40.0286748
T           -0.06427241  0.3443322
```

Which of the three models above (A, B or C) have unbiased estimates of β_1 if model (1) is correct?

Each of the models have unbiased estimates of β_1.

(c) Which of the three models above (A, B or C) have valid confidence intervals for β_1 if model (1) is correct? If a model has invalid confidence intervals, what is wrong with the confidence interval?

While the three models all have unbiased estimates of β_1 they compute standard errors differently. As we know the variances are proportional to $1/N$, the correct weighting is model B. The other models will have valid centers for their confidence intervals, but the width will be incorrect as the standard errors were incorrect.