KEY

STAT 411 FINAL EXAM

December 14, 2017, 198 pts

1. **TRUE/FALSE**: (18 pts, 3 pts each) For each of the problems below, circle **T** for True or **F** for False.

   (a) **T** / F Increasing the confidence level produces a wider confidence interval.

   (b) **T** / F Confidence intervals generated by regression are robust to deviations from the constant variance assumption.

   (c) **T** / F In the population of humans, type of diet is correlated with weight loss.

   (d) **T** / F When deciding between two models, smaller VIF values suggest which model is "better".

   (e) **T** / F When deciding between two models, smaller AIC values suggest which model is "better".

   (f) **T** / F When assessing case influence statistics, *The Sleuth* recommends starting with a basic model with only essential interactions.

2. (16 pts) State the Central Limit Theorem and explain how it helps scientists perform the scientific method.

If $X_1, \ldots, X_n$ are independent with mean μ and SD $\sigma \neq 0$ and if n is large, then $\bar{X} \dot{\sim} N(\mu, \sigma^2/n)$. This substantiates the use of z, t, and F tests when testing hypotheses.

3. (16 pts) When using CIs to estimate the mean number of avalanches in the Absaroka Mountains as a function of the angle of the slope that holds the snow, when should you use Workman-Hotelling CIs instead of individual CIs?

W-H 95% CIs of mean number of avalanches maintains 95% confidence over a potentially large family of CIs, one CI for each angle for lots of angles. An individual CI is appropriate for the mean number of avalanches for only a few angles.

4. A high school ski coach is interested in estimating the effect of temperature on her skiers' times to complete a 1.5km course at Bridger Bowl. She recruits 15 Intermediate skiers and 15 expert skiers between the ages of 16 and 20 from Bozeman High to participate in a small study. Each skier is timed (in seconds) and the temperature (in Fahrenheit) is recorded. R code and partial output:

```
m1 = lm(time ~ Ability * temp)
summary(m1)
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
##(Intercept)               263.099     18.652  14.106 1.08e-13
##AbilityIntermediate       -64.074     23.799  -2.692  0.01225
##temp                        1.274      1.440   0.885  0.38443
##AbilityIntermediate:temp    7.454      1.661   4.489  0.00013
##
##Residual standard error: 24.93 on 26 degrees of freedom
##Multiple R-squared:  0.858,     Adjusted R-squared:  0.8416
##F-statistic: 52.37 on 3 and 26 DF,  p-value: 3.721e-11
```

(a) (8 pts) Write out the regression model that includes all relevant parameters.

$\mu\{\text{time} \mid \text{Ability}, \text{temp}\} = \beta_0 + \beta_1 \text{Dummy}_{\text{Int}}(\text{Ability}) + \beta_2 \text{temp} + \beta_3 \text{Dummy}_{\text{Int}}(\text{Ability}) \times \text{temp}.$

(b) (6 pts) Why is it important to include skiers with different ability levels?

Because expert skiers are faster than others and this may confound the times as a function of temp if not accounted for.

(c) (16 pts) Do the data suggest that the effect of temperature on ski times is dependent on skiers' ability levels? Provide the estimate(s) of the effect, p-value(s), and interpret in terms of the problem using non-technical language.

Yes - intermediate skiers' times increase by 7.5sec / 1°F increase more than the expert skiers ($p = 0.00013$). Also the times of intermediate skiers are 64sec. faster (?) than expert skiers at 0°F ($p = 0.012$)

(d) (16 pts) The last line of the regression output above allows you to compare two models. What are the two models being compared? Write out the appropriate null and alternative hypotheses, give the value of the test statistic, report the p-value, and state a conclusion.

The null model ($\mu = \beta_0$) is being compared to regression model in 4(a). $H_0: \beta_1 = \beta_2 = \beta_3 = 0$; $F = 52.37$; $p = 3.721 \times 10^{-11}$

$H_a: \beta_j \neq 0$ for some j.

∴ Data suggest that at least one of the coefficients ($\beta_1 \to \beta_3$) non-zero; i.e. model in 4a is better than null model.

(e) (8 pts) Give the estimated slope for Intermediate skiers and interpret in terms of the problem (including units).

$\hat{\beta}_2 + \hat{\beta}_3 = 1.274 + 7.454 = 8.728$

(f) (6 pts) Explain how you would you calculate the standard error of the estimated slope for Intermediate skiers.

$SE(\hat{\beta}_2 + \hat{\beta}_3)^2 = SE(\hat{\beta}_2)^2 + SE(\hat{\beta}_3) + 2\,Cov(\hat{\beta}_2, \hat{\beta}_2)$

$= 1.44^2 + 7.45^2 + 2(\;\;)$

This covariance from $\hat{\sigma}^2(X^TX)$ matrix

(g) (8 pts) Calculate the numerical value of the sum of the squares of the residuals. **SHOW YOUR WORK!**

Residual SE = 24.93 $\Rightarrow$ MSF = $24.93^2 = 621.5 = \frac{SSF}{DFF}$

$\Rightarrow$ SSF = MSF × DFF = 621.5 × 26

= 16159.13

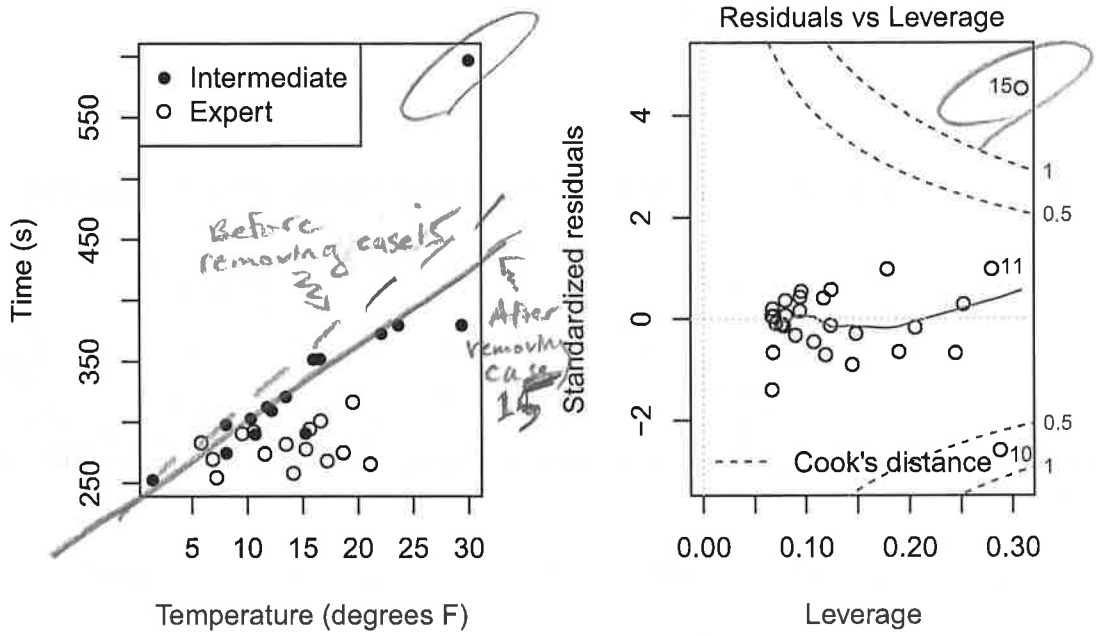
(h) (8 pts) Do the data suggest which group of skiers is faster at 0 degrees F? Intermediate or Expert skiers? How much faster? Report the p-value.

Yes, Intermediate skiers are faster by 64s (!!)
($p = 0.012$)

(i) (8 pts) A colleague wants you to drop the interaction from the model to make the model and interpretations simpler. In R, you type in `m2=lm(time ~ Ability + temp)` then perform a test via `anova(m2,m1)`. The p-value associated with this test is provided in the regression output on the previous page (yes, it is). Report the p-value. Does this result support your colleague's opinion to drop the interaction? Explain.

The extra sum of squares test gives $p = 0.00013$ which suggests that more Residual SS are explained by a model that includes the interaction. I.e., the model with interaction in 4c is better than model without an interaction.

(j) The left pane in the next plot shows the skiers' times for different temperatures; the right pane displays case influence statistics for the regression model fit to these data.



i. (6 pts) Which unusual case(s) show up as having the potential to influence the regression? Circle the unusual case(s) in the left AND right panes above.

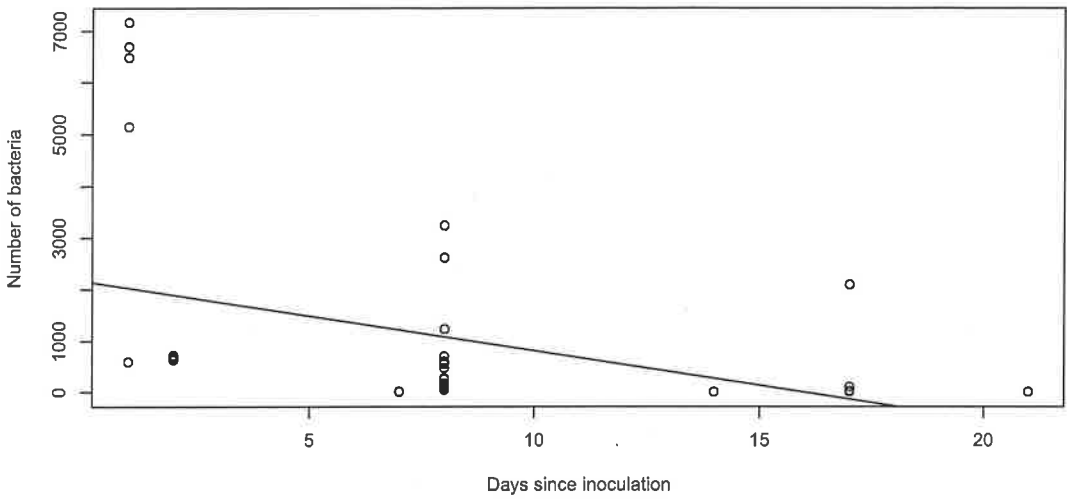
ii. (8 pts) In practice, you would remove the unusual case(s) that you identified above and re-fit the regression model to determine the influence of the unusual case(s). Of the 4 regression coefficient estimates in the R output on a previous page, use the pane on the left to guess which of the 4 estimates will be influenced and how the estimates will be influenced (i.e., which estimates go UP, which estimates go down?).

$\hat{\beta}_1$ will INCREASE (because y-int for intermediate will INCREASE)

$\hat{\beta}_3$ will DECREASE (because slope for intermediate will DECREASE)

(k) (18 pts) You decide to permanently remove the unusual case(s) from the data set and report the updated regression output in a paper you are about to submit to a journal. State a *Scope of Inference*.

the highschool skiers may likely be representative of skiers from BZN HS, so these results only pertain to BZN HS skiers between 16 and 20. This was an observational study, so we may not conclude that increased temperatures caused the skiers times to increase. Because we dropped case 15, the results only pertain to times <400s and to temps <30°F.



5. At a microbiology laboratory, bacteria were inoculated onto $n = 68$ pennies. Each penny was sampled once at some time point over 3 weeks and the number of bacteria on the penny was recorded. The data and the *least squares regression line* are plotted above.

   (a) (8 pts) What is meant by *least squares* and what does it have to do with the regression line in the figure?

This means that the line gives the smallest sum squares of the residuals compared to any other line fit to the data.

   (b) (16 pts) What 4 assumptions are required for any inference to be valid? Use the scatterplot to assess 2 of the 4 assumptions.

Independence: this was an experiment, in a controlled laboratory environment, so independence of bacteria on different pennies seems reasonable.

Linearity - This appears violated, instead →.

Constant Variance - This appears violated, with variance decreasing as Days increase.

Normality - cannot assess, need normal probability plot.

   (c) (8 pts) Should the data be transformed? If so, explain why, indicate what transform you would try and which variable(s) you would transform.

The non-linearity (→) and decreasing variance suggest log(Num. bacteria) only.