Remember, I want your name only on the back of the last page of your answers. If you have answers written on that page, please put your name on a clean sheet of paper.

R and SAS code and output will be found after the questions. Unless specifically stated, all R models use the default contrasts (contr.treatment for lm and contr.helmert for lmer).

1. A consulting client describes to you their study. They want to compare growth rates of pigs. The treatments are all possible combinations of 5 diets and whether or not (2 levels) the pig is injected with monensin, an antibiotic. The experiment will be conducted in a barn with 15 pens. Each pen will have 4 pigs. Pens are randomly assigned to diet (3 pens per diet) and pigs are randomly assigned to monensin or not (2 pigs per pen per trt). The study will be repeated at a total of 3 locations. In the entire study, there are 3 locations, 45 pens, and 180 pigs. The mean growth rate is expected to vary among locations. The effect of the antibiotic (monensin) is expected to be more consistent across the locations than is the effect of diet.

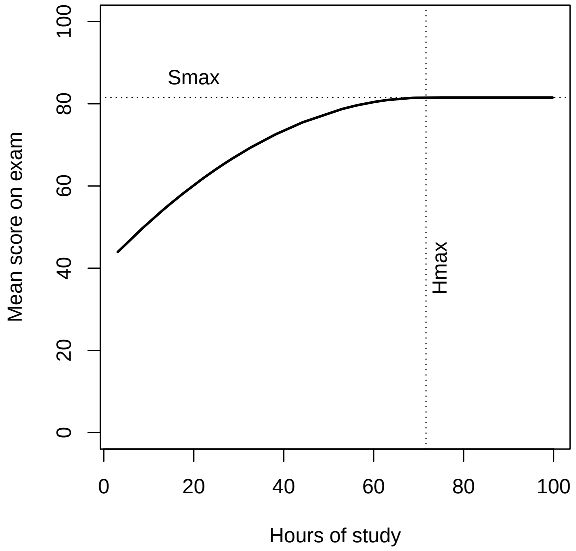
   (a) 10 pts. Write out the skeleton ANOVA table for this study. List the sources of variation and their degrees of freedom.

   (b) 5 pts. This study will be used to make recommendations for pig producers at new locations. Which sources of variation should be considered random? (Only need to list the random ones)

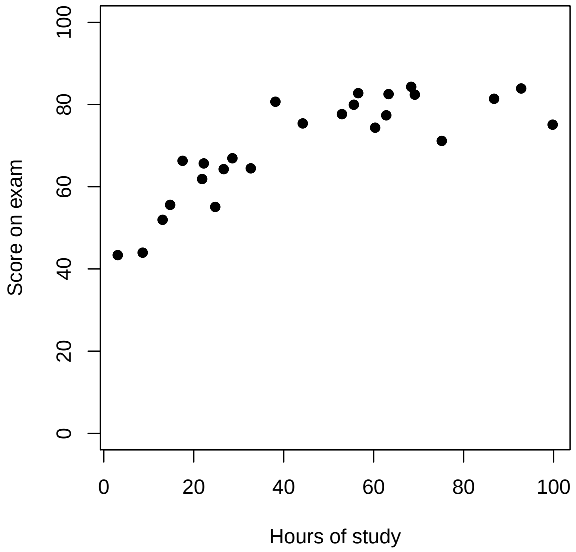
   (c) 5 pts. Based on what you know of the study and its goals, what is the most appropriate error term (denominator) for the F test of the diet main effect?

2. The data for this question are made up based on a study of the relationship between the length of time studying for a final exam and the score on that final. The presumption is that more studying will increase the student's score. The investigators believe there is a quadratic relationship between the number of study hours and the final score, but that past a certain number of hours more studying does not help. For simplicity, we will call that maximum number of hours the "max hours".

   The relationship between # hours studying and mean score is shown below. I have indicated values of S_{max} and H_{max} on the plot.



You have data from 25 students studying for and taking a math final exam. The data are plotted here:



(a) 5 pts. Write an appropriate model for the relationship between # hours studying and the student's score. If you can not do this part, ask and I will give you the model.

R code and output for this model are on pages 1 and 4 of the R packet. SAS code and output are on pages 1 and 3 of the SAS packet. In the R code, quadtop() is a function that models the relationship between # hours studying and the mean score.

(b) 5 pts. What are the LS estimates of S_{max}, H_{max}, and the quadratic coefficient, b_2? Be sure to report answers with an appropriate number of digits (and not more).

(c) 5 pts. Calculate the Wald 95% two-sided confidence interval for H_{max}. Some potentially useful T and Z quantiles are $T_{0.975,25} = 2.059$, $T_{0.975,22} = 2.074$, $T_{0.95,25} = 1.708$, $T_{0.95,22} = 1.717$, $Z_{0.95} = 1.645$, and $Z_{0.975} = 1.960$.

(d) 5 pts. The profile 95% confidence interval for H_{max} in the R output is (42.1, 114.4). This is somewhat different from the Wald interval. Explain why these two intervals are not the same.

(e) 5 pts. You are asked to report one confidence interval in a paper. Which would you choose, the Wald interval or the profile interval? Briefly defend/explain your choice.

3. You decide to evaluate fit of the non-linear regression by comparing that fit to the fit of a penalized regression spline.

   R code and output are on pages 1 and 4 of the R packet. There is no SAS.

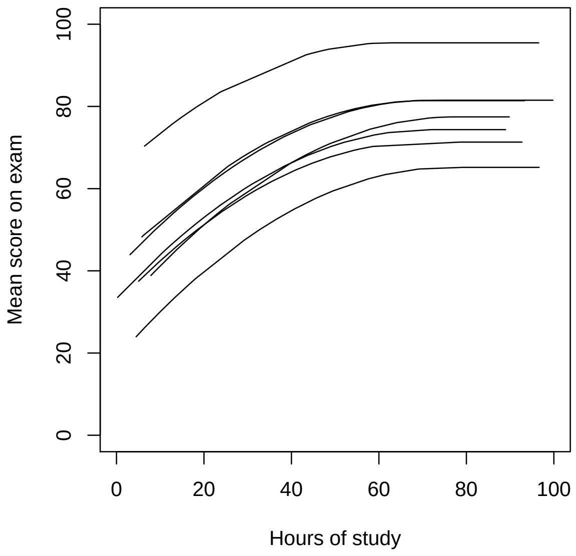
   (a) Calculate the error df for this penalized spline fit. If you can't calculate this, use 21.2 as the error df.

   (b) You first decide to test whether there is any association between hours spent studying and test score. You do this by comparing a model with just an intercept to the fit of the penalized spline. The SS Error for the intercept-only model is 3718.1. Calculate the F statistic for this test and give its distribution under the null hypothesis.

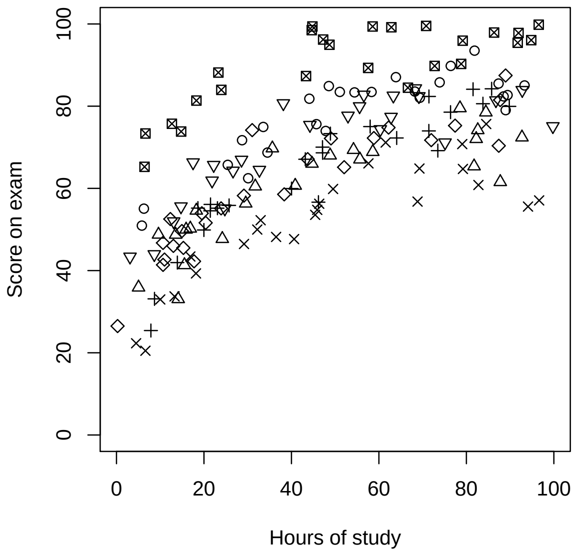
   (c) Calculate the model comparison F statistic comparing the penalized spline to the non-linear regression fit in question 2b. The error SS for the nls fit is 480.82. Report the apparent F statistic (it should look unusual).

   (d) The apparent F statistic is negative. Explain why the model comparison in the previous question is an invalid comparison.

4. The study of effort and test score was actually done in seven different subjects. The investigators expect that the maximum score and the "max hours" vary by subject, but that the curvature (quadratic coefficient) is the same for all subjects. A plot of the expected relationship between hours studying and mean score is shown below. Each line represents one subject.



The investigators collect data on 25 students in each of 7 subject areas (math, statistics, english, french, history, biology and chemistry). Different students were tested in each subject area(so there are 175 rows of data from 175 students). The variability in student scores around the mean score is anticipated to be normally distributed with an unknown but constant variance. A plot of the data is shown below. In this plot, different symbols indicate different subject areas.



(a) 10 pts. The investigators want to make general conclusions about all subject areas, so they choose to model the maximum score, S_{max_i}, and "max hours", H_{max_i} as random quantities, varying by subject area. The quadratic coefficient should be considered the same value for all subject areas. Write an appropriate model for the data from all 7 subject areas. Define all subscripts and give distributions for all random quantities on the right-hand side of the model equation (use the usual choices).

If you can not do this part, ask and I will give you the model.

(b) 5 pts. How many random effects are in this model? Briefly explain your answer. (Hint: If this were a linear mixed effect model, the answer would be the number of columns of the $\boldsymbol{Z}$ matrix for your model in question 4a.)

(c) 5 pts. I actually considered three models for the random effects:

Model	VC matrix for the random effects	AIC	BIC	lnL
nl1	$[\sigma_S^2]$	1214.2	1230.0	-602.09
nl2	$\begin{bmatrix} \sigma_S^2 & \sigma_{SH} \\ \sigma_{SH} & \sigma_H^2 \end{bmatrix}$	1104.5	1126.7	-545.26
nl2b	$\begin{bmatrix} \sigma_S^2 & 0 \\ 0 & \sigma_H^2 \end{bmatrix}$	1103.8	1122.8	-545.90

Construct a test of $\sigma_{SH} = 0$, i.e., no covariance between $Smax_i$ and $Hmax_i$. Report your test statistic and give its distribution under Ho.

(d) 5 pts. Construct a test of $\sigma_H^2 = 0$. Report your test statistic and give its distribution under Ho.

Rightly or wrongly, you decide to use model nl2b for future analysis. R code and output are on pages 1-2 and 5-6 of the R packet. There is no SAS output (integration issues).

(f) 5 pts. Predict the score for a new statistics student who studied for 80 hours.

(g) 5 pts. Predict the score for a new physics student who studied for 30 hours.

5. The cbpp data are a classic data set on the incidence of contagious bovine pleuropneumonia (CPBB) in zebu cattle in Ethiopia. The Binomial response is the number of new cases of CPBB occuring in a herd over a 3 month period. Data were collected on 15 herds for 4 periods. The variables are:

herd	factor identifying the herd (1 to 15)
incidence	# new cases of CBPP for that herd and period
size	# cattle in the herd
period	factor identifying the period (1 to 4)

The analysis will be based on the model below, or extensions or simplifications of it:

$$\begin{aligned} Y_{ij} &\sim \text{Binomial}(N_{ij}, \pi_{ij}) \\ \text{logit } \pi_{ij} &= \mu + \alpha_i + \beta_j \\ \beta_j &\sim N(0, \sigma^2), \end{aligned}$$

where:

Y_{ij}:	is the # of new cases in herd i and period j,
N_{ij}:	is the herd size for herd i and period j,
π_{ij}:	is the probability of a new case in period i and herd j,
$\mu + \alpha_i$:	mean logit for period i,
β_j:	random effect for herd j.

There are 56 observations in the data set. One herd was only measured in period 1; one was only measured in periods 1-3. R code and output is on pages 2-3 and 6-9; SAS code and output is on pages 1-2 and 4-8. The R models used the lm default contr.treatment. The SAS models used 'set last to 0' parameterization.

(a) 5 pts. You consider 3 models for random effects:

cbpp.m0	No random effects, β_j omitted
cbpp.m1	Herd random, β_j included
cbpp.m2	Herd random, and an additional random effect for each observation to model potential overdispersion

Which model for random effects do you consider most appropriate? Briefly justify your choice.

(b) 5 pts. No matter what model you chose in the previous part, use cbpp.m1 (only Herd random) as the random effects model for subsequent parts. The output also includes output for the cbpp.m10 model, in which period is removed. Construct a test of the null hypothesis that all periods have the same incidence. Report your test statistic and give its distribution under the null hypothesis.

(c) 5 pts. Estimate the subject-specific log-odds ratio comparing incidence in period 2 to that in period 3.

(d) 5 pts. The Ethopian Department of Animal Health wants to estimate the average incidence in period 1 in the entire country. Consider herds to be a random sample of all herds in the country. If possible with the information available, provide them that estimate.

Note: The back transformation of the logit function is $\pi = \frac{1}{1+exp(-\text{logodds})}$

There are 10 points for free. Good luck on the summer exams!