Final Exam Instructions

- The exam and dataset will be made available to you at 9:00am.
- The exam is due electronically before 12:00 midnight on the day after you receive the exam. LATE EXAMS WILL BE PENALIZED 10% FOR EACH HOUR LATE!
- The exam contains short answer questions and a report question. Your solutions to the short answer questions can be either hand written or typed.
- Your solution to the report question should be in report form. Type your report using LaTeX or Word or other professional software. Save your report as a pdf file. Separate your report into sections, as appropriate.
- Turn in your exam in 3 files. Replace "Yourname" with your first and last name:
  1. "YournameShortAnswer.pdf": a pdf of your solutions to the short answer questions (this may be a scan of handwritten work).
  2. "YournameReport.pdf": a pdf of your solution to the report question.
  3. "YournameRcode.r": a text file containing the R-code you used for your analysis of the report question. Do NOT include the R output. Only include the R commands run. Include enough comments that someone could reproduce your work if needed.
- You may NOT consult with any other person in any way, electronically or in person, about this exam.
- Clearly cite any sources used in your analysis.
- Turning in this exam indicates that you have read and agree with the following statement:

  I affirm that the report I have submitted is my work, and my work only, and that I have neither given nor received any help or information from any other person. I have cited any written sources I used to complete my report. I recognize that giving or receiving any help on this exam constitutes cheating and will result in my receiving a zero grade on this exam, as well as potentially other consequences under PSU's Policy on Academic Integrity.

Short Answer. Answer the following questions. These short questions are completely unrelated to the report question on the following page. Your answers here may be typed or they may be hand-written, and your answers to these questions do NOT count towards the 10-page limit of the report.

1. Assume that you observe (x_i, y_i) for $i = 1, 2, \ldots, n$ and you know that the data arise according to the following model:

$$y_i = k_i + \frac{1}{2a} x_i^2, \quad k_i \sim N(b, cx_i^2)$$

where each k_i is independent of all others, and a, b, and c are constants to be estimated. Explain how you could estimate a, b, and c, if such estimation is possible. If it is not possible to estimate some or all of these parameters, clearly explain why not.

2. Consider the Gaussian linear model

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I})$$

where $\mathbf{y}$ is a $n \times 1$ vector of observations and $\mathbf{X}$ is an $n \times p$ matrix of predictor variables (including an intercept).

Assume you have reason to believe that the first three observations $\{y_1, y_2, y_3\}$ may jointly come from a different distribution than the rest of the data. Construct a test with null hypothesis that these three observations come from the specified linear regression model and alternative hypothesis that they are outliers. Clearly describe your test statistic, any parameter estimates used in the test statistic, the distribution of the test statistic, and how you would reject the null hypothesis.

Report. Conduct an analysis of the data described below. Write your analysis in report format. Describe the stated problem of interest and give a brief introduction to the data, including any important exploratory data analysis. Clearly describe any models or methods used in your analysis, and justify your use of the models. Interpret the results of your analysis. Use technical statistical terminology. Include any relevant figures and tables. Label all figures, and clearly describe the importance of the figure in the text of your report. Do not include any figures in your report that are not described in the text of your report. Close your report by summarizing your important findings and discussing the strengths and weaknesses of your analysis.

Your report must be no more than 10 pages long, including plots, tables, and all description. To keep to this page limit, you may need to carefully choose what to include in the report. You may mention briefly analyses without including the results in the report. For example, instead of showing multiple residual plots, you might choose to state the residual analyses you conducted and briefly describe the results. Turn in an r-script file containing ALL CODE used in your analysis, with enough comments that someone could reproduce your analysis if they wanted to.

House Loss to Wildfires. Severe wildfires in southern California in 2009 present a rare opportunity to study which factors influence whether a house near a forest is burned in a wildfire, or is not. The "fire.Rdata" file contains data from 487 randomly selected houses in neighborhoods judged to be "at risk" to forest fires. Many characteristics of these houses were obtained from satellite images and other remote geospatial data sources. The state of California is interested in using this data to make recommendations to homeowners on how to maintain their property to minimize the risk of fire damage to their home.

The variables collected for each house in "fire.Rdata" are described on the next page of this document. In your report, you should focus on interpreting the effect of the 6 "characteristics near the home" (planted, buildings, perc.woody, perc.cleared, distance2bush, distance2tree), as these are items that homeowners have control over. All other variables are considered to be potentially related to burn status, but are not the main focus of your analysis.

Focus your report and analysis on the following questions.

1. Clearly describe the relationship between the 6 "characteristics near the home" (planted, buildings, perc.woody, perc.cleared, distance2bush, distance2tree) and the probability that a home is burned in a wildfire. Provide clear descriptions of any important relationships, and plots if they would be helpful. Of particular interest is the potentially different fire risk of houses with similar nearby tree characteristics (perc.woody and distance2tree) but different "planted" status. The "planted" variable is coded as "r" if the nearby vegetation is a remnant of the surrounding forest (the vegetation existed before the house was built), or coded as "p" if the nearby vegetation was planted by homeowners.

2. Compare multiple approaches for estimating regression parameters to obtain the best model you can for predicting which houses will burn and which will not. What percent of the time is your prediction of burn status correct in your best predictive model?

3. Use your best predictive model to compare the effects that each of the following three proposed management practices would have on house loss due to wildfire.

   (a) Remove all trees within 10 meters of all houses.

   (b) Require any home with at least 50% woody vegetation within 40m of the house to remove vegetation until they only have 50% woody vegetation within 40m of their house.

   (c) Replant any "remnant" vegetation within 40m of all houses with identical "planted" vegetation.

   The state of California is planning to launch a campaign to promote one of these three actions, and wants to determine which of these three practices is likely to be the most beneficial in preventing house loss to wildfire. Provide an estimate of the number of houses that would have been burned in the 2009 fires if one of the three proposed management actions had been taken before 2009. Which of the three proposed management actions will be most beneficial?

Variable	Description
Burnt	Indicator variable coded as "1" if the house was burnt and "0" if the house was not burnt.
	Variables related to the house's location
slope	Calculated in degrees. Higher values indicate homes on steeper terrain.
ffdi	Forest Fire Danger Index. A measure of how susceptible to fire a house is. See: Noble IR, Bary BAV, Gill AM (1980) McArthur's fire-danger meters expressed as equations. Australian Journal of Ecology 5: 201–203.
aspect	Calculated as degrees from north.
topo	Topographic Position. Calculated at each house as one of seven levels (1=ridge to 7=valley bottom)
Edge	Distance (m) to forest edge.
	Characteristics Near the Home
planted	A visual assessment of whether woody vegetation within a circle with a radius of 40m from the centroid of each house was predominantly planted ("p") or remnant ("r") using the pre-fire imagery.
buildings	The number of buildings within a radius of 40m from the centroid of each house.
perc.woody	% mapped woody vegetation within a circle with a radius of 40m from the house
perc.cleared	Visual estimate of % of land without woody vegetation within a circle with a radius of 40m from the centroid of each house using the pre-fire imagery.
distance2bush	Distance from each house to nearest bush
distance2tree	Distance from each house to nearest tree or shrub (m) visible.
	Characteristics of the Transect (straight line) to the boundary of the 2009 fire in the upwind directiod
adj.for.type	Categorical variable. The nearest mapped forest type structure (open forest, woodland, non-forest) to each house in the upwind direction.
amt.cleared	Amount (m) of open space calculated along a transect in the upwind direction from each house to the 2009 wildfire boundary.
amt.woody	Amount (m) of mapped woody vegetation calculated along a transect in the upwind direction from each house to the 2009 wildfire boundary.
amt.burntless5yrs	Amount (m) of land from each house that was burnt ≤5 years prior to 2009.
perc.burntless5yrs	% of landscape burnt ≤5 years ago calculated along a transect from each house to the 2009 wildfire boundary in the upwind direction .
amt.not.burnt5to10yrs	Amount (m) of land from each house that was not burnt for >5-≤10 years prior to 2009 measured in the upwind direction.
perc.burnt5to10yrs	% of landscape burnt >5-≤10 years ago calculated along a transect from each house to the 2009 wildfire boundary in the upwind direction
amt.unlogged	Amount (m) of land from each house that is unlogged in the previous 30 years in the upwind direction.
perc.logged	% of landscape logged in the 30 years prior to 2009 calculated along a transect from each house to the 2009 wildfire boundary in the upwind direction.
amt.not.NP	Amount (m) of land from each house that is not National Park in the upwind direction.
amt.not.SF	Amount (m) of land from each house that is not State Forest in the upwind direction.