1. Mesquite plants

(a) The assumption for our regression model is that the **errors** are normally distributed; not the dependent variable. Thus there is no problem with the observation of a non-normal marginal distribution of total weight of harvestable leaves with all other variables in the regression held fixed.

One other important point: the central limit theorem would argue that $\hat{\beta}$ has a normal distribution even if the errors do not have a normal distribution; the CLT does not “fix” the non-normality of the errors.

(b) A one unit (i.e., one meter) increase in canopy diameter (short axis) is associated with a 309.3 gram increase in the expected total weight of harvestable leaves with all other variables in the regression held fixed.

(c) The relevant t-critical value is the two-sided one, $t_{40.975} = 2.021$, which tells us that the intercept, canopy1 and density are significantly different from zero at the .05 level.

(d) There is not enough information provided to test this hypothesis. There are two approaches that would work: (1) find s.e.$(\hat{\beta}_{c1} - \hat{\beta}_{c2})$ and then do a t-test using $\hat{\beta}_{c1} - \hat{\beta}_{c2}$ divided by its standard error; (2) run a “reduced” model in which the two variables are assumed to have the same coefficient (this would mean including their sum in the regression instead of the two separate predictors) and then do an F-test comparing this reduced model to the full model.

(e) The normal probability plot shows some evidence of non-normality. There is one (or perhaps a few) big positive residuals and some big negative residuals as well. The plot of standardized residuals versus fitted values definitely suggests increasing variance. It could suggest non-linearity but that’s likely just a reflection of the one big outlier.

(f) I apologize because this question was not very clearly worded. My intention was for you to both report on the relevant diagnostics and on the data values that may be associated with the unusual diagnostic. A full answer would mention cutoffs for leverage ($3x6/46 = .39$), Cook’s distance ($50$th percentile of $F_{6.46} = .90$), and externally studentized residuals ($t_{39.975} = 2.02$ or Bonferroni adjusted version $t_{39,1-.(5/2460)} = 3.53$). It would also discuss the cases: case 3 has high leverage and high (but not unusually high) Cook’s distance. This seems to be a result of the unusual canopy height compared to total height (very different than the rest of the data); case 27 is not terribly unusual but does have density 5 (which is the 3rd highest value); case 28 is clearly unusual on all-dimensions and highly influential - it is much larger and denser than any other tree; case 35 is also unusual with a large residual, large leverage and large Cook’s distance - though no diagnostic surpasses its threshold, the combination and the high value of density mark this as an unusual case; case 40 is not terribly unusual - it has a standardized residual larger than 2 in absolute value but we would expect a couple of such values in a data set this size. The main summary here is that on this scale we may want to restrict model building to plants with density 3 or below and collect more data to predict larger plants. On the other hand the logarithmic transformation seems to solve much of this problem. In terms of grading, I took a few points off if you didn’t say anything about data or if you didn’t say anything about the diagnostic measures. Finally, I would like to emphasize the important point that every case with standardized residual bigger than 2 in absolute value is NOT an outlier - we expect some values of this size.

(g) My intention was for you to consider what the transformations imply about the relationship of the variables on the original scale. The log/log model is multiplicative with coefficients corresponding to exponents of the independent variables, i.e., $\log(Y_i) = \beta_0 + \beta_1 \log(X_{1i}) + \beta_2 \log(X_{2i}) + \epsilon_i$ leads to $Y_i = \exp(\beta_0)X_{1i}^{\beta_1}X_{2i}^{\beta_2}\exp(\epsilon_i)$.

(h) This question required you to think about the relationship of the confidence interval for the mean response at a specified value and the prediction interval for a new response at a specified value. You are given information about the former. The CI is obtained as estimate plus/minus 2.021*(standard error of the predicted mean) which means that we could solve for the standard error of the predicted mean as $(5.8419-5.6946)/2.021 = .0729$. Then the standard error for the prediction is $\sqrt{MSE + (s.e.predicted.mean)^2} = \sqrt{1617 + .0729^2} = .4087$. This gives a prediction interval on the log scale of 5.6946 +/- 2.021*.4087 = (4.8687, 6.5205). And exponentiating gives a prediction interval on the scale of total weight equal to (130.2, 678.9).

2. Cloud seeding

(a) Model 1

i. For treatment 2 we find that $\text{RAIN} = (2.3044 - 1.0423) + (.4538 + .7934)\text{PRED} = 1.2621 + 1.2472\text{PRED}$.  

ii. This can happen when predictors are highly correlated. This tends to happen with interactions because the interaction is correlated with each element that makes up the interaction, e.g., the interaction of treat1 and pred is highly correlated with treat1. Many people just restated the question noting that these are different tests; I was looking for more of a “why”.

(b) Model 3 assumes that the relationship between LOG(RAIN) and LOG(PRED) is linear for each group with the same slope but different intercepts. This is the ANCOVA or parallel lines model. Model 4 fits a single line (same intercept, same slope for all treatments).
(c) $H_0 : \beta_{\text{treat,LOGPRED}} = \beta_{\text{treat2,LOGPRED}} = 0$ vs $H_a : \text{not } H_0$.

Use $F = ((\text{SSE}(\text{model3}) - \text{SSE}(\text{model2}))/2)/\text{MSE}(\text{model2}) = (14.41757 - 13.59675)/2/25179 = 1.63$. Compare to $F_{2,25}$ distn and find that $P$-value $>.10$. Do not reject $H_0$. Model 3 appears to be adequate.

(d) It is important to learn how to describe the $P$-value. Please make sure you understand. The $P$-value measures the probability of observing a test statistic as big or bigger than the observed value (here a t-statistic more extreme than 1.8) if the null hypothesis is true.

(e) Key point is that because it is a randomized study the three groups (control and two treatments) should each have similar distributions of predicted rainfall. This means you will get the approximately the same treatment effect estimates with or without LOGPRED in the model. It is likely that the inference will be more precise with LOGPRED in the model (smaller MSE and narrow CIs) but the inference for treatment effects is still valid when LOGPRED is omitted.

(f) To assess whether the randomization was appropriate you should focus on PRED not RAIN. The latter includes the s.d.s; the s.d.s are relevant but not the most important aspect. For full credit you needed to commented on the s.d.s. For others. Then the regression version of the model is

$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$ with $\mu$ the grand or overall mean, $\alpha_i$ the effect of water type $i$, $\beta_j$ the effect of curing time $j$, $\gamma_{ij}$ the interaction effect of water type $i$ and curing time $j$, and $\epsilon_{ijk}$ the normal mean zero, variance $\sigma^2$ error term.

(g) $R^2$ never decreases when we add a predictor; it can stay the same or increase. As a result it tends to favor big models and will often lead us to overfit to the specifics of our data.

(h) To validate the model, we would use Model 2 to predict outcomes in the new (unseen) data set. We would then compare the average squared prediction error in the new data set to the MSE(model2) = .25. If the average squared prediction error is larger then it suggests Model 2 is not reliable. For some reason, many people talked a lot about model comparison here but the focus was on evaluating one model using the test set.

3. Mortar production

(a) The ANOVA model is usually written as $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$ with $\mu$ the grand or overall mean, $\alpha_i$ the effect of water type $i$, $\beta_j$ the effect of curing time $j$, $\gamma_{ij}$ the interaction effect of water type $i$ and curing time $j$, and $\epsilon_{ijk}$ the normal mean zero, variance $\sigma^2$ error term.

(b) To implement as a regression model we need to define indicator variables. For example, you can let $W_1 = 1$ for hard water, $0$ for soft water; $C7_i = 1$ for curing time of seven days, $0$ for others; and $C28_i = 1$ for curing time of 28 days, $0$ for others. Then the regression version of the model is $Y_i = \beta_0 + \beta_1 W_i + \beta_2 C7_i + \beta_3 C28_i + \beta_4 W_i \ast C7_i + \beta_5 W_i \ast C28_i + \epsilon_i$. This can be written in matrix form with the matrix $X$ consisting of a column of all 1’s (corresponding to the intercept), then columns for the three indicators defined above, and finally two columns for the interactions (products of the relevant indicators). The $\beta$ coefficients are interpreted as $\beta_0 = \text{reference group}$ (soft water, 3 days) mean, $\beta_1$ measures the effect of hard water compared to soft water for the 3 day group, $\beta_2$ measures the effect of 7 day curing vs 3 day curing for the soft water condition, $\beta_3$ measure the effect of 28 day curing vs 3 day curing for the soft water condition, $\beta_4$ measures the difference of the 7-day vs 3-day comparison between hard water and soft water, and $\beta_5$ measures the difference of the 28-day vs 3-day comparison between hard water and soft water. You didn’t have to say all of this for full credit but did need to define appropriate indicators that matched your model in (a).

(c) ANOVA table

i. The MSE = 8880/48 = 185. This is an estimate of $\sigma^2$, the variance among repeated observations with a given curing time and water type. It was not sufficient to give the formula and mention that this measures the average difference between $Y_{ijk}$ and $\bar{Y}_{ij}$ because it is of more interest to us as an estimate of the model/population parameter (not as a summary statistic for these data).

ii. $F_{\text{water}} = 864/185 = 4.67; F_{\text{curing}} = 1300.5/185 = 7.03; F_{\text{inter.}} = 166.5/185 = 0.90$. The critical value for the water type test is $F_{1,48,.95} = 4.04$ and the critical value for the curing time and interaction tests is $F_{2,48,.95} = 3.19$. Thus water type and curing time are statistically significant at the .05 level.

(d) Contrast

i. Define $\gamma = \mu_{h,28} + \mu_{s,28} - 0.5\mu_{h,7} - 0.5\mu_{s,7} - 0.5\mu_{h,3} - 0.5\mu_{s,3}$ to be the contrast of interest. (I should have divided those weights in half to help with the interpretation in the next part.) Then to get the confidence interval we calculate $\hat{\gamma} = (59+58-0.5(56+44+36+47)) = 25.5$ and s.e.($\hat{\gamma}$) = $\sqrt{185*(0.5^2 + 0.5^2 + 0.5^2 + 0.5^2 + 1^2 + 1^2)}/9 = 7.85$. Then the confidence interval is $25.5 \pm 2.021 \ast 7.85 = (9.64, 41.36)$.

ii. We are 95% confident that (twice) the difference between average tensile strength when curing for 28 days and average tensile strength when curing for 3/7 days is between 9.6 and 41.4 tensile strength points.

(e) Regression would allow for predictions at curing times other than those used in the experiment. Regression however requires a fairly strong linear assumption and would not yield useful results if the relationship were nonlinear.
Data Analysis - Lung Function

To grade the data analysis I identified a number of criteria that I looked for. This included some issues associate with the writeup and some issues associated with the analysis.

- Quality of the writeup / report - Should be a thorough discussion with judicious use of tables/figures. Several reports include lots of output with a few sentences of text added in. This is very difficult to read. Only relevant pieces of the output should be included with the text of the report. Also, you don’t have to mention obvious things such as the fact that ID was not a helpful variable.

- Assessment of a basic regression model - Almost everyone tried an initial regression model that led to a transformation of the response. This was very good.

- Avoid irrelevant statistical analyses - You are told that age/height are important variables when it comes to lung function. It doesn’t make sense to spend much time analyzing FEV vs smoking status. Any such analysis will be impacted by the confounding effects of age. A simple observation that smokers appear to have higher FEV is interesting; a formal test is not relevant.

- Diagnostics / outliers - It is important that you carry out such an examination. Here one point was more influential than the others but none of the Cook’s distance values indicated that there would be a substantial change by deleting the observation. There were a number of high leverage points – I don’t have a good explanation for this it seems to be the fact that there are relatively few smokers. Again, none of the high leverage points end up being influential. There were a couple of very large studentized residuals but these don’t seem to be obvious errors so better to leave them in. In the end I did not see any reason to delete observations.

- Model selection - I viewed the use of formal model selection tools here (all subsets regression) as being unnecessary. There are a small enough number of predictors that you should be able to navigate the analysis without trying all possible models.

- Interactions - The question emphasized determining whether "the effect of smoking is purely additive, or whether it depends on the level of other factors". This required consideration of whether there should be interactions. Most solutions did not find interactions helpful. Failure to investigate was viewed as a substantial negative. If you did include interactions, then it is crucial that you interpret them carefully.

- Summary - The report summary is important. It should not just report on significance. It should emphasize the magnitude of the effect and also provide some degree of uncertainty around the estimate (i.e., a confidence interval)