Take Home Portion

These problems are to be done independently. You may use your notes, your textbook, and the notes I have provided for this class. And, of course, you may use R for statistical analysis. But you may not consult any other people or materials. If you have questions, send me an email.

Due date: 1pm, Tuesday, May 8

In each problem below a statistical study is outlined and you are asked to analyse the data. Your analysis should include both R output and your interpretation of it (in sentence/paragraph form). Prepare your analysis using Word or some other software that allows you to copy and paste R output and submit a hard copy of your solutions. See the information included with your project assignments for more details about getting graphs out of R into your solutions.

- Be sure to comment on the appropriateness of your analysis to the data and any transformations you use.
- For R output, include the R command as well as the output in your solution. You may edit the output to remove extra information that you did not use and gets in the way. (I have often done this on the class handouts.)

1. Thirteen specimens of 90/10 Cu-Ni alloys with varying percents of iron content were submerged in sea water for 60 days. After 60 days, the weight loss due to corrosion was recorded in units of milligrams per square decimeter per day. The data from this study can be obtained using `getData('m243/corrosion') -> rust`. The values of Fe are percents, so 1.5 means 1.5% iron in the alloy.
   a) Does the iron content in the alloy matter?
   b) Predict the amount of corrosion in an alloy with 1.75% iron. (Include both a point estimate and an interval estimate).
   c) Predict the amount of corrosion in an alloy with 3.0% iron. (Include both a point estimate and an interval estimate).
   d) Are there any reasons to worry about either of these predictions?

2. Patients with advanced cancers of the stomach, bronchus, colon, ovary or breast were treated with ascorbate. The purpose of the study was to determine if the survival times differ with respect to the organ affected by the cancer. The data can be obtained using `getData('m243/cancer') -> cancer`. Analyze the data and answer the following questions
   a) Is ascorbate equally effective in all cancers tested?
   b) If not, comment on the differences.
   c) List at least one reason other than the cancer’s response to the drug that could explain differences in survival times from organ to organ. (Do this even if these data do not suggest there is a difference.)
In Class Portion

Show all your work. Unsupported answers will not receive full credit. **No mystery numbers please!** If you use R or your calculator, be sure to indicate both what you entered and the result. If you want to load the R extras I have written, type `getCalvin('m243')`.

3. **True/False.** Regression is sensitive to outliers.

4. A regression line is sometimes called a “least squares line”. Explain where this name comes from.

The regression can be defined as the line that minimizes the sum of the squares of the residuals (vertical distance between the line and the data points).

5. Side-by-side boxplots comparing three treatment groups in each of three studies are shown below. Assuming the sample sizes are the same in each study, which one will have the smallest p-value when testing for equality of means?

6. If the 95% confidence interval for a mean is $3.2 \pm 2.8$, what can you say about the p-value of the hypothesis test with $H_0 : \mu = 0$, and $H_a : \mu > 0$.
   
   a) $P > 0.10$,
   b) $0.05 < P < 0.10$,
   c) $0.025 < P < 0.05$,
   d) $P < 0.025$,
   e) It is impossible to tell which of these is the case from the information given.
What do I do? In each of the following situations, pretend you want to know some information and you are designing a statistical study to find out about it. Give the following THREE pieces of information for each: (i) what variables you would need to have in your data set (ii) whether they are categorical or quantitative, and (iii) what statistical procedure you would use to analyze the results.

Select your procedures from the following list: 1-proportion, 2-proportion, 1-sample Z, 1-sample t, Paired t, 2-sample t, 1-way ANOVA, 2-way ANOVA, simple linear regression, logistic regression, multiple regression. (Some procedures may be repeated or omitted. For some situations there may be more than one solution – one correct solution suffices.)

7. You want to know if grade school students perform better on timed arithmetic tests when there is music playing in the background or when the room is quiet.

Variables:

Procedure:

8. You want to know if the systolic blood pressure (big number part of blood pressure reading) of patients who arrive at the emergency room in a coma is related to whether or not they survive.

Variables:

Procedure:

9. You want to compare the amount of fat in Twinkies, Hohos, and Oreo cookies. (These are all sweet snack foods.)

Variables:

Procedure:

10. You want to investigate the relationship between high school GPAs and college GPAs for students at the University of Michigan.

Variables:

Procedure:
11. The critical flicker frequency (cff) is the highest frequency (in cycles/sec) at which a person can detect the flicker in a flickering light source. At frequencies above the cff, the light source appears to be continuous even though it is actually flickering. An investigation carried out to see if average cff depends on iris color yielded the following results:

<table>
<thead>
<tr>
<th>Eye Color</th>
<th>Blue</th>
<th>Brown</th>
<th>Green</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>28.167</td>
<td>25.588</td>
<td>26.920</td>
</tr>
<tr>
<td>st deviation</td>
<td>1.5280</td>
<td>1.3653</td>
<td>1.8431</td>
</tr>
<tr>
<td>number of subjects</td>
<td>6</td>
<td>8</td>
<td>5</td>
</tr>
</tbody>
</table>

a) Supply the missing values in the following ANOVA table.

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>0.023</td>
<td></td>
<td></td>
<td>0.023</td>
</tr>
<tr>
<td>Residuals</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>61.3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

d) List two underlying assumptions of the ANOVA test used here.

c) What is the conclusion of the test?

e) One of these assumptions can be “checked” using output on this page, which one? What do you conclude?

f) What does the output below tell us? Explain.

Tukey multiple comparisons of means
95% family-wise confidence level

Fit: aov(formula = cff ~ color)

$color
diff lwr upr
Brown-Blue -2.5792 -4.73550 -0.42284
Green-Blue -1.2467 -3.66440 1.17106
Green-Brown 1.3325 -0.94372 3.60872
12. A study was done to measure the effect of various characteristics of cheddar cheese on its perceived taste. The amounts of lactic acid and hydrogen sulfide in the cheese and the average taste score from a number of taste testers were recorded. Here is some output from R fitting two models.

**Model 1**

```r
> summary(lm(taste~Lactic+H2S,cheese))

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -21.23e+01 | 1.121e+01  | -1.894  | 0.06903  |
| Lactic         | 3.028e+01  | 8.012e+00  | 3.780   | 0.00079  |
| H2S            | 7.723e-04  | 4.223e-04  | 1.829   | 0.07853  |
```

Residual standard error: 11.28 on 27 degrees of freedom  
Multiple R-Squared: 0.5515, Adjusted R-squared: 0.5183  
F-statistic: 16.6 on 2 and 27 DF, p-value: 1.99e-05

**Model 2**

```r
> summary(lm(taste~Lactic+log(H2S),cheese))

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -27.592  | 8.982      | -3.072  | 0.00481  |
| Lactic         | 19.887   | 7.959      | 2.499   | 0.01885  |
| log(H2S)       | 3.946    | 1.136      | 3.475   | 0.00174  |
```

Residual standard error: 9.942 on 27 degrees of freedom  
Multiple R-Squared: 0.6517, Adjusted R-squared: 0.6259  
F-statistic: 25.26 on 2 and 27 DF, p-value: 6.551e-07

---

a) Which model do you like better? Be specific about how you made your selection.

b) What taste score does each model predict for cheese with \( \text{lactic} = 2 \) and \( \text{H2S} = 200 \). (No interval required.)

c) Give a 95% confidence interval for the coefficient on \( \text{lactic} \) in Model 2.
1. A simple linear regression is a reasonable choice. It can be only modestly improved by a transformation, and I accepted models both with and without a transformation. A scatter plot indicates a strong negative association between iron content and corrosion. Furthermore, the p-value for the model utility test (with either transformed or untransformed model) is very small, and $r^2$ is quite large. These all indicate the usefulness of iron content for estimating corrosion. Most of you had no trouble using `predict()` to make an interval prediction. Do make sure you understand the difference between a confidence interval and a prediction interval. The prediction for a 3% iron content alloy is less reliable because we have no data in that range, so we have no way to know if the pattern we see continues. Unless we have some theoretical reason to expect a certain type of relationship to continue, we must always worry about such extrapolations. In addition, prediction intervals are very sensitive to the shape of the population distribution. So if we use a prediction interval, we need to be convinced that it is reasonable to assume that the distribution of loss is approximately normal for each level of iron content.

2. This analysis requires a transformation. Many of you did not even check if the assumptions of ANOVA seemed reasonable. The variances in the different groups are quite different from one another, and the residuals do not look normal (based on a normal quantile plot). Remember that you can get useful pictures using the `plot()` command, or you can generate your own using `resid()` to get the residuals. A log transformation of the survival time produces a model that fits the assumptions much better. The p-value for the resulting ANOVA test is small enough to reject the null hypothesis that the mean (log) survival time is the same for each cancer when treated with ascorbate. Following up with Tukey’s pairwise comparisons shows which cancers differ significantly from which. The results you get here from the log-transformed model and an untransformed model don’t match exactly, but in both cases we see that breast cancer has significantly longer survival time than some of the other cancers. This could of course be due to things other than that cancer’s (better) response to the drug. Perhaps, for example, the survival times differ even if one doesn’t take the drug – perhaps because of different diagnosis patterns, age of onset, etc.

3. True. A single observation that does not fit the pattern of the rest of the data can change the regression line a lot, especially if the value of the explanatory variable is quite large or quite small compared to the others.

5. The p-value will be smallest in data set B because the group means are more different than they are in data set A (where the variances are comparable) and the variances are smaller than in data set C (where the group means are comparable).

6. d. $P < 0.025$. Since 0 is not in the 95% CI, we know that the two-sided hypothesis test would be significant at the $\alpha = 0.05$ level. Thus a one-sided test will have a p-value that is less than 0.25.

7. background noise (cat); test performance (quant); two-sample t
   Or test performance with music (quant); test performance without music (quant); paired t

8. blood pressure (quant); survival (cat); logistic regression

9. fat (quant), type of snack (cat); 1-way ANOVA

10. high school GPA (quant); university GPA (quant); simple linear regression
11. We can compute the sum of squares for color (SSTr) be remembering that

\[ SSTr = \sum (\text{group mean} - \text{grand mean})^2, \]

and calculating the grand mean as

\[
\text{grand mean} = \bar{x} = (6 \cdot 28.167 + 8 \cdot 25.588 + 5 \cdot 26.920)/(6 + 8 + 5) = 26.75295.
\]

This gives

\[
SSTr = 6 \cdot (28.167 - \bar{x})^2 + 8 \cdot (25.588 - \bar{x})^2 + 5 \cdot (26.920 - \bar{x})^2 = 22.99362
\]

The rest of the table is then straightforward:

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F</th>
<th>value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>2</td>
<td>23.0</td>
<td>11.5</td>
<td>4.8</td>
<td>0.023</td>
</tr>
<tr>
<td>Residuals</td>
<td>16</td>
<td>38.3</td>
<td>2.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>18</td>
<td>61.3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

12. a) Model 2 is better. Normality of residuals is at least as good as for Model 1, the variation of the residuals about 0 is more consistent across the range of fitted values (although this is a bit hard to judge in model 1 since there are fewer observations with the large fits), and most importantly \( r^2 \) is higher, so Model 2 is able to explain a larger proportion of the variance in the data with the same number of parameters.

b) Model 1: \(-21.2 + 30.3 \times 2 + .00077 \times 200 \equiv 39.5\)
Model 2: \(-27.59 + 19.887 \times 2 + 3.946 \times \log(200) \equiv 33.09\)

c) \(19.887 \pm t \cdot \hat{\sigma} = 19.887 \pm 2.051 \cdot \hat{\sigma} = 19.887 \pm 16.33\)