Problem 1 (10 points)

A study of area high school students shows that the average high school GPA of students who have never smoked is 0.4 points higher than the average high school GPA of students who smoke at least six cigarettes per day. A p-value in a one-sided t-test of the hypothesis of no difference is 0.049. The investigators conclude that there is statistically significant evidence that smoking causes a decrease in GPA in high school students. Comment on the appropriateness of this conclusion using statistical reasoning.

Solution: The explanatory variable is observational, as students make their own choice about whether or not to smoke. Thus, it is not justified to attribute causation. There is slight evidence of a difference between groups, but this could be caused by smoking or any number of other confounding factors.

Problem 2 (10 points)

In a designed experiment in which each subject is given one of two treatments decided at random, the p-value in a hypothesis test for no difference in treatment means is 0.06. The investigators conclude that there is a 6% chance that the population means are identical. Briefly comment on the validity of this conclusion.

Solution: A p-value is the probability of obtaining a test statistic at least as extreme as that actually observed. It is not the probability that the null hypotheses is true.

Problem 3 (10 points)

Consider a scatter plot of two variables with a simple linear regression line added to the plot. For each narrow vertical strip, the points look to be fairly evenly distributed above and below the line in a roughly symmetric fashion with similar amounts of variability around the line in each strip. However, the x values are strongly skewed to the right— there are many points at low x values and relatively few at high x values. Is a simple linear regression adequate, or should you consider a transformation of the x values so that their distribution would be more approximately normal?

Solution: In simple linear regression, we do not assume that the explanatory variable follows a normal distribution. The description of the plot indicates that the regression fits well. There is no apparent need to consider a transformation of the x values, but we could check to see if some points are unduly influential.

Problem 4 (20 points)

Twenty epileptic patients participated in a study of a new anticonvulsant drug, valproate. The subjects were randomly divided into two groups of ten. The first group received valproate for eight weeks during which the number of epileptic seizures was counted. The second group received a placebo and was observed at the same time. Afterwards, the treatments were switched for everyone and they were again observed for eight weeks and counts of epileptic seizures were recorded. The variables `Placebo` and `Valproate` record for each subject the number of seizures under each treatment.

Here is the output from two possible analyses of the data.

```
> t.test(Placebo, Valproate, paired = T)

        Paired t-test

data:  Placebo and Valproate
t = 3.9586, df = 19, p-value = 0.0008422
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  5.890863 19.109137
sample estimates:
mean of the differences
                   12.5

> t.test(Placebo, Valproate, var.equal = T)

        Two Sample t-test
```

```
data:  Placebo and Valproate
t = 2.3758, df = 38, p-value = 0.02266
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  1.848795 23.151205
sample estimates:
mean of x mean of y
     29.2      16.7
```

(a) Intepret the results of the most appropriate analysis (both the test and the confidence interval) in the context of the problem.

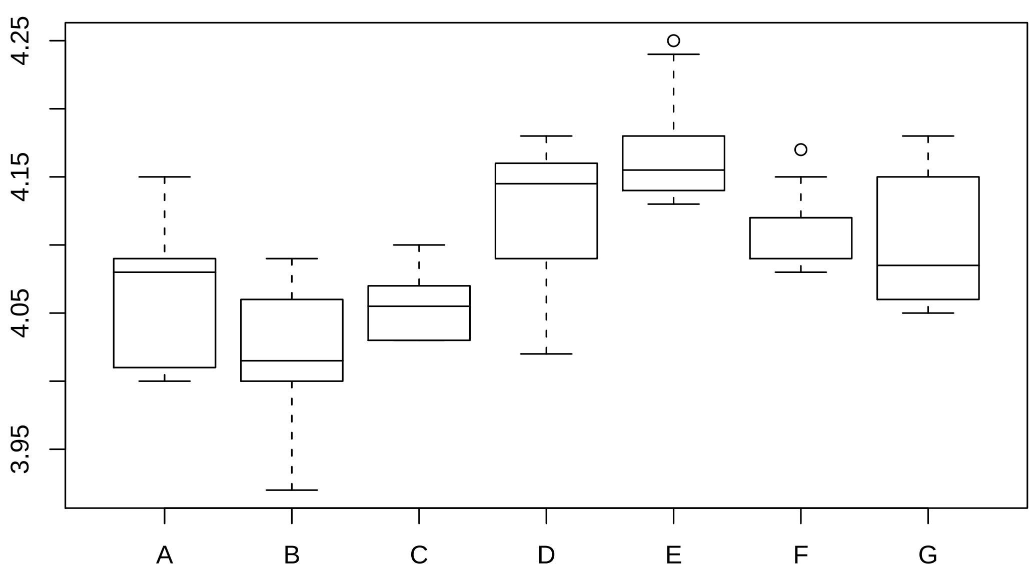
(b) State one assumption made by the inferential methods which you interpreted in the previous part and describe a plot you could make to examine the validity of this assumption.

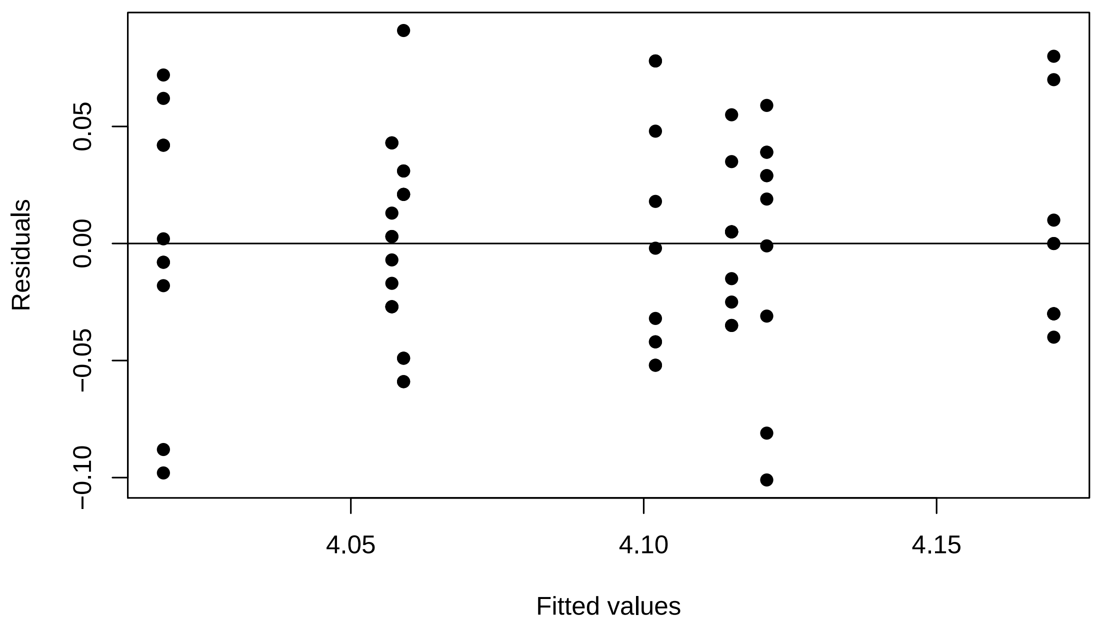
Solution:

(a) This is a paired design, so a paired test is most appropriate. There is very strong evidence that valproate caused a decrease in the number of epileptic seizures among these subjects ($t = 3.96$, one-sided $p = 0.0004$ in a paired t-test). We are 95% confident that the mean decrease is seizures in this time period would be between 5.9 and 19.1. This result may not generalize to other subjects.

(b) The method assumes that the differences are normally distributed (although it is robust to non-normality). A boxplot of individual differences would indicate if the results are affected by skewness or outliers.

Plots for Problem 5





Problem 5 (30 points)

A manufacturer of medicinal tablets was interested in testing the measurement accuracy of seven different laboratories. A composite was prepared by grinding and mixing together many tablets. Each lab made ten separate measurements on tablet-sized portions of the composite that had nominal dosage levels of 4g per tablet of the active ingredient.. The table below shows the mean and standard deviation of these measurements (in grams) of the active ingredient from each lab. The previous page shows side-by-side boxplots of the data and a residual plot after fitting a model in which each sample has its own mean.

	A	B	C	D	E	F	G
mean	4.059	4.018	4.057	4.121	4.170	4.115	4.102
sd	0.051	0.058	0.027	0.054	0.043	0.029	0.052

(a) Complete the partial ANOVA table.

```
Analysis of Variance Table

Response: weight
                Df       Sum Sq      Mean Sq     F value    Pr(>F)
-------------------------------------------------------------------
lab                    0.153434                             6.975e-09

Residuals              0.135360
-------------------------------------------------------------------
Total
```

(b) List the main assumptions of ANOVA. For each assumption you list, indicate if the plots and statistics above provide evidence that the assumption might not hold.

(c) State the null and alternative hypotheses of the ANOVA. If you introduce statistical notation, define what it means. (For example, don't use μ_1 without saying what this represents in the context of the problem.)

(d) Summarize and interpret the results of the ANOVA in the context of the problem. If you are able to infer a causal relationship, do so. Indicate the scope of your inferences — how can they be generalized? Briefly discuss how your answer to part (a) relates to your interpretation of the results.

(e) Suppose that you were interested in estimating the difference in the population means for laboratories A and B. Calculate the margin of error for a 95% confidence interval for this difference. (If you cannot find numerical values for everything you need, indicate qualitatively what you would need to complete the computation.)

(f) Suppose that you were further interested in testing for differences between each pair of population means. Pick a method that adjusts for multiple comparisons and justify your selection. (Do not, however, actually carry out the tests.)

Solution:

(a) Complete the partial ANOVA table.

```
Analysis of Variance Table

Response: weight
              Df      Sum Sq      Mean Sq     F value    Pr(>F)
-----------------------------------------------------------------
lab            6    0.153434    0.025572     11.90      6.975e-09

Residuals     63    0.135360    0.00214857
-----------------------------------------------------------------
Total
```

(b) Assumptions —

- Normal Distributions: residuals look fairly symmetric without extreme values
- Equal Variance: the sample standard deviaitons are the same order of magnitude, but some groups have SDs about twice as large as others.
- Independent Samples: the plots don't address this assumption.

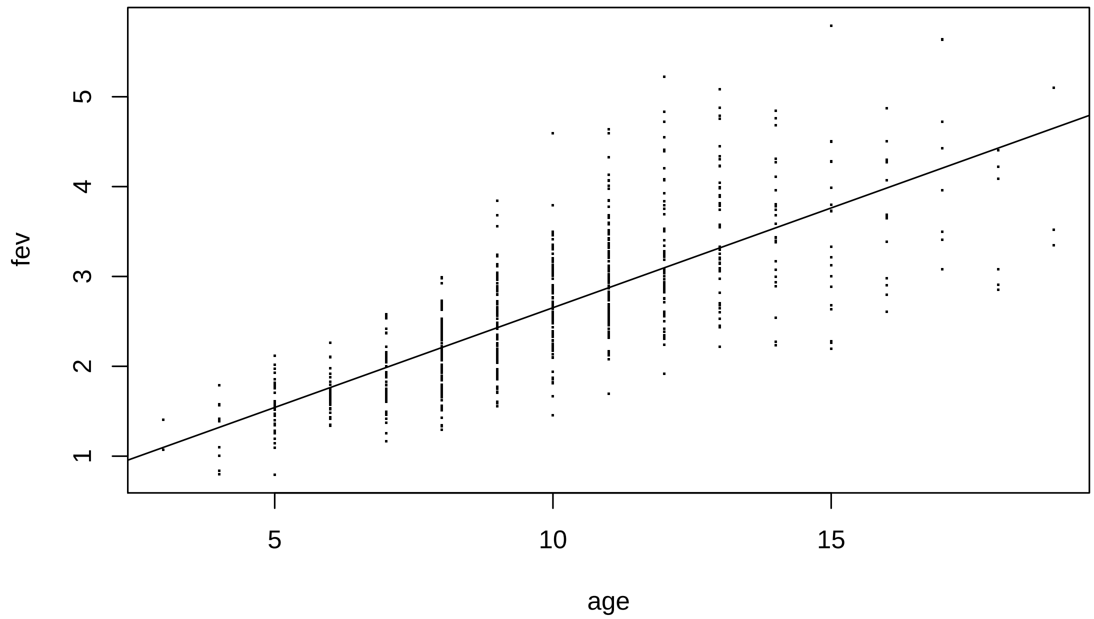
(c) $H_0 : \mu_A = \mu_B = \cdots = \mu_G$ versus H_A : not all means are equal, where μ_X is the mean of measurements in lab X for this type of measurement.

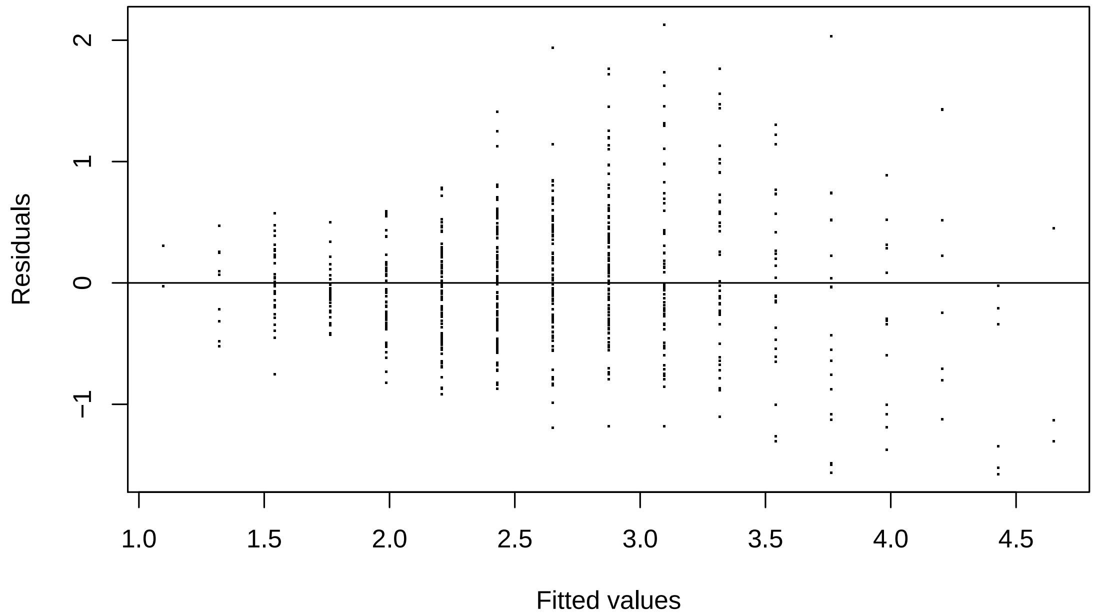
(d) There is strong evidence that the labs have different means for their measurement procedures. Because the true mass of the active ingredient should have been very close to equal (as a result of the grinding and mixing process), some labs must have biased procedures. These results would not generalize to other labs.

(e) Margin of error $= t\hat{\sigma}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. Plugging in numbers gives $1.998 \times \sqrt{0.00214857} \times \sqrt{\frac{1}{10} + \frac{1}{10}} \doteq 0.041$

(f) The Tukey-Kramer procedure would be a good choice because it is designed specifically for the examing all pairwise comparisons between means when sample sizes are equal.

Plots for Problem 6





Problem 6 (20 points) Forced expiratory volume (FEV) is a measure of the volume of air that a person can exhale in one second when blowing with maximum force. FEV tends to increase as children get older and taller. A scatter plot of FEV versus age for several hundred children with a linear regression line fit and a plot of the residuals are shown on the previous page. A summary of the regression line is below. What features do you notice in the plots? What would be your next step of an analysis? Briefly explain.

```
> summary(fit)

Call:
lm(formula = fev ~ age)

Residuals:
     Min       1Q   Median       3Q      Max
-1.57539 -0.34567 -0.04989  0.32124  2.12786

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.431648   0.077895   5.541 4.36e-08 ***
age         0.222041   0.007518  29.533  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5675 on 652 degrees of freedom
Multiple R-Squared: 0.5722,       Adjusted R-squared: 0.5716
F-statistic: 872.2 on 1 and 652 DF,  p-value: < 2.2e-16
```

Solution: The residuals tend to get larger as the fitted values increase. There is also evidence of non-linearity. Transformation to the response variable (by logs, for example) is a good thing to try first.