BIOS 662, Fall 2018

Homework 2

Assigned: Tuesday, September 4

Due: Tuesday, September 11

1. In the "Datasets" sub-folder of the "Homework materials" folder under "Resources" on the Sakai web site for this course, there is a dataset "HW2_SBP.txt". The data are systolic blood pressures of 40 women who had an MI (myocardial infarction, i.e. heart attack) less than two years after their blood pressure was measured and of 160 women who did not have an MI within two years. Each row in the dataset corresponds to one woman. The first observation in a row is 1 for those who had an MI within two years and 0 otherwise. The second observation is the systolic blood pressure.
   (a) Use R or SAS to draw a histogram and boxplot of systolic blood pressure for all 200 women (that is, do not separate those who did and did not have an MI).
   (b) Using the definition of percentile from the class notes, compute the 25th, 50th (i.e., median) and 75th percentiles.
   (c) Determine the IQR.
   (d) Find the largest observation $\leq 75^{\text{th}}$ percentile + 1.5 IQR and the smallest observation $\geq 25^{\text{th}}$ percentile $-$ 1.5 IQR (i.e., the extent of the "whiskers"). Based on these results, does the computed boxplot appear to agree with the definition of a boxplot from our notes? If not, investigate the discrepancy and report your findings.
   (e) Use a plot to compare the distribution of systolic blood pressure in those who had an MI against that of those who did not. Do blood pressures in the two groups appear to differ? If so, in what direction?

2. The dataset "HW2_PGE.txt" in the "Datasets" sub-folder contains the data in Table 3.20 of the textbook. The last value in each row is 1 for the patients with hypercalcemia and 0 otherwise.
   (a) Obtain the mean and standard deviation of plasma iPGE separately for patients with and without hypercalcemia. Do you think there is enough evidence to conclude that the means of the two groups differ? (Later in this course we will study more formal ways to compare the two means.)
   (b) Do part (c) of Problem 3.15 of the textbook.
   (c) The values for one patient appear to be particularly anomalous. Identify this patient. Suppose it was determined that there had been an

error in measuring the patient's serum calcium. Suggest a value for serum calcium that would be more consistent with the patient's plasma iPGE value and the pattern in the rest of the data.

(d) Without re-doing any of your calculations, what effect do you think changing the serum calcium value to the one you suggested would have on the means and standard deviations in the first part of this problem?

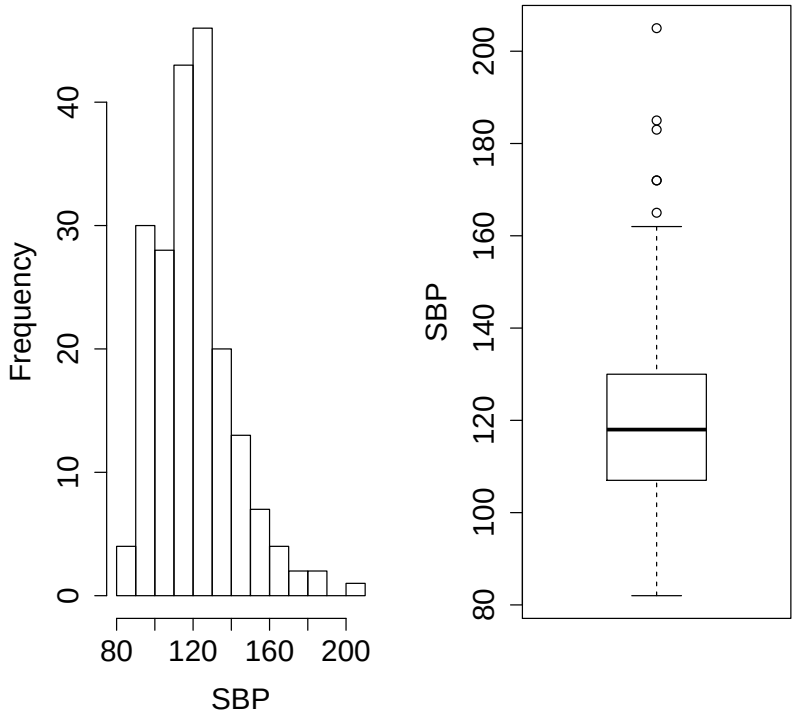
BIOS 662

Homework 2 Solution

September, 2018

Question 1

(a) Using the R functions hist() and boxplot(), we get the following output:



(b) Because $n = 200$ and $p = 0.25$, $np = 50$ is an integer, so the 25th percentile is given by

$$\hat{\zeta}_{0.25} = \frac{y_{(50)} + y_{(51)}}{2} = \frac{107 + 107}{2} = 107$$

Similarly, one can show $\hat{\zeta}_{0.5} = 118$ and $\hat{\zeta}_{0.75} = 130$.

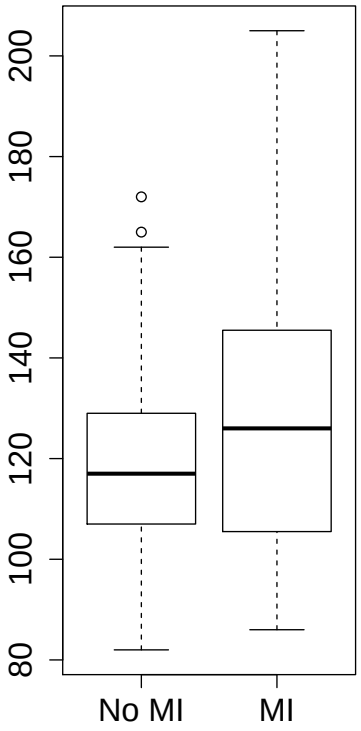
(c) Thus the IQR equals $130 - 107 = 23$.

(d) The 75th percentile + 1.5 IQR $= 130 + 1.5 \times 23 = 164.5$ and the largest observation less than this is 162. Likewise, the 25th percentile $-$ 1.5 IQR $= 107 - 1.5 \times 23 = 72.5$ and the smallest observation greater than this is 82. Looking at the histogram we see that there are no outliers atthe lower end of the distribution, which is why in the boxplot there

are no individual observations plotted below the lower whisker. The results agree with R exactly:

```
> boxplot(sbpall)$stats
     [,1]
[1,]   82
[2,]  107
[3,]  118
[4,]  130
[5,]  162
```

(e) The following figure shows side-by-side boxplots for the two groups. SBP tends to be higher in the MI group, with its median almost as high as the third quartile of the no MI group and its upper whisker extending substantially beyond the largest SBP in the no MI group.



Question 2

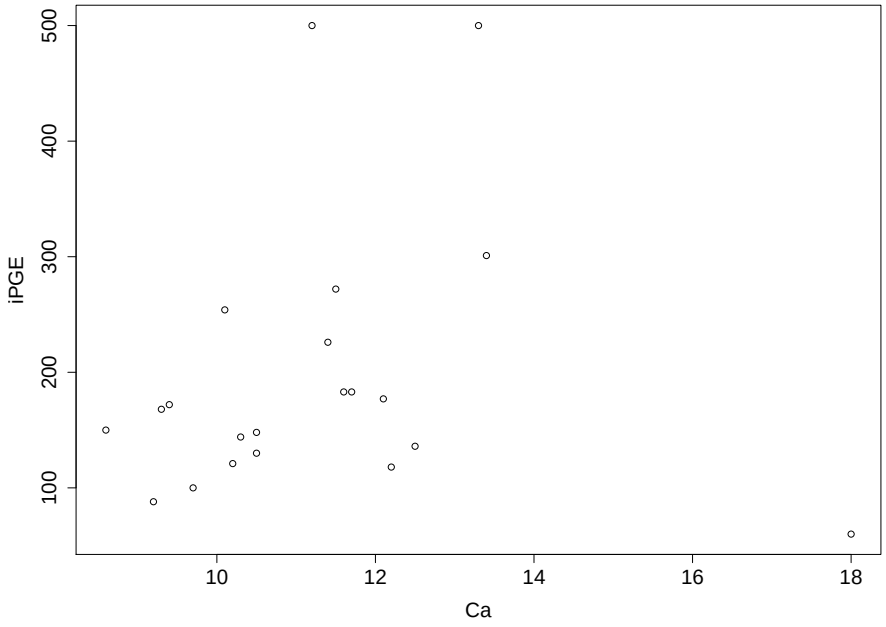
(a) $\bar{X}_{\text{Hypercalcemia}} = 2656/11 = 241.5$; $\bar{X}_{\text{Normocalcemia}} = 1475/10 = 147.5$.

$s^2_{\text{Hypercalcemia}} = \frac{1}{11-1}\left(849988 - 11 \cdot 241.5^2\right) = 20868.47$; so $s = 144.46$.

$s^2_{\text{Normocalcemia}} = \frac{1}{10-1}\left(236749 - 10 \cdot 147.5^2\right) = 2131.83$; so $s = 46.17$.

The means seem to be substantially different. We'll see in the coming weeks that we need to use information about the standard deviations in order to decide whether the corresponding population means really do appear to differ.

(b) Below is a scatterplot of plasma iPGE against serum Ca.



If we ignore a few outliers, there is some evidence that higher plasma iPGE levels tend to be associated with higher serum Ca levels. There is substantial variability from person to person though, so if person A has higher serum Ca than person B it does not automatically follow that person A will have higher plasma iPGE than person B.

(c) Patient #11 has the highest serum Ca value among all patients yet has the lowest plasma iPGE level. A serum calcium level below 10 would be more in keeping with the tendency for lower plasma iPGE to be associated with lower serum Ca.

(d) Patients with serum calcium above 10.5 mg/dL are classified as hypercalcemic. If the serum calcium value for patient #11 is really below 10, then this patient would be classified as not having hypercalcemia and so would be moved from one group to the other. Because this patient has plasma iPGE level well below all the others in the

Hypercalcemia group, moving this patient out of the group would result in a larger mean plasma iPGE value for the group. By removing a value far from the mean, the standard deviation will decrease. The patient's plasma iPGE level is also lower than all values in the other group, so moving the patient into that group would lower its mean plasma iPGE value and because the newly added value is more extreme than other values in the group, the standard deviation will increase.

Suppose we change the serum calcium value for patient #11 from 18 to 10. Then the sample means and standard deviations of plasma iPGE change from

$\bar{X}_{\text{Hypercalcemia}} = 241.5; \quad \bar{X}_{\text{Normocalcemia}} = 147.5.$

$s_{\text{Hypercalcemia}} = 144.46; \quad s_{\text{Normocalcemia}} = 46.17$

to

$\bar{X}_{\text{Hypercalcemia}} = 259.6; \quad \bar{X}_{\text{Normocalcemia}} = 139.5.$

$s_{\text{Hypercalcemia}} = 138.43; \quad s_{\text{Normocalcemia}} = 57.13.$

BIOS 662, Fall 2018

Homework 3

Assigned: Tuesday, September 18

Due: Tuesday, September 25

1. This is a continuation of problem #2 from Homework 2 involving the dataset "HW2_PGE.txt". Use the data for all patients combined without regard to hypercalcemia status.
   (a) Use a plot to decide whether the distribution of serum calcium is approximately normal.
   (b) Calculate the sample mean and standard deviation for serum calcium, and construct a 95% confidence interval for the population mean serum calcium of such patients.
   (c) Suppose that the sample size is doubled (but yielding the same mean and standard deviation). Determine the percentage change in the width of the 95% confidence interval. Repeat assuming a sample three times the original size.
   (d) Use the bootstrap method to obtain a 95% confidence interval for the population mean serum calcium of such patients.
   (e) Calculate the sample median and obtain an exact 95% confidence interval for the population median serum calcium of such patients.

2. Problem 4.20 on page 111 of the textbook.

BIOS 662

Homework 3 Solution

September, 2018

Question 1

(a) A QQ plot of serum calcium is given on page 3. Most of the points lie close to a straight line but the point in the top right corner is way off the line. Recall that in part 2(c) of Homework 2 it was suggested that one of the calcium values was recorded incorrectly. It is the anomalous value that is in the top right corner. If this value is changed from 18 to 7.5 we get the second of the QQ plots on page 3. In that one all the points lie reasonably close to the line. Later in the semester we will look at a more formal test of normality.

For the rest of the question we will use the uncorrected calcium value.

(b) The population variance is unknown, so we have to handle this either as a small sample from the normal distribution (with unknown variance) or a "large" sample from an unknown distribution. Here $n = 21$, $\bar{Y} = 11.27$, $s^2 = 4.164$ and $s = 2.041$.

If this is a small sample from the normal distribution, a 95% CI for μ is:

$$\bar{Y} \pm t_{n-1,1-\alpha/2}(s/\sqrt{n}) = 11.27 \pm 2.086 \times 2.041/\sqrt{21} = (10.34,\ 12.20).$$

If this is a large sample from an arbitrary distribution, using the Central Limit Theorem and Slutsky's Theorem a 95% CI for μ is:

$$\bar{Y} \pm z_{1-\alpha/2}(s/\sqrt{n}) = 11.27 \pm 1.96 \times 2.041/\sqrt{21} = (10.40,\ 12.14).$$

The sample is a little too small to qualify as a large sample and if we leave the value of 18 uncorrected, the normality assumption is also questionable, so neither CI is very satisfactory here.

(c) If in (b) we assume the sample is from the normal distribution, the width of the confidence interval is

$$2 \times t_{n-1,1-\alpha/2}(s/\sqrt{n}) = 2 \times 2.086 \times 2.041/\sqrt{21} = 1.86.$$

Doubling the sample size changes the degrees of freedom of t, so $t_{n-1,1-\alpha/2} = t_{41,0.975} = 2.02$ and the width of the confidence interval is

$$2 \times t_{n-1,1-\alpha/2}(s/\sqrt{n}) = 2 \times 2.02 \times 2.041/\sqrt{42} = 1.27.$$

This is a reduction of $100 \times (1.86 - 1.27)/1.86 = 31.5\%$.

Tripling the sample size, $t_{n-1,1-\alpha/2} = t_{62,0.975} = 2.00$ and the width of the confidence interval is

$$2 \times t_{n-1,1-\alpha/2}(s/\sqrt{n}) = 2 \times 2.00 \times 2.041/\sqrt{63} = 1.03.$$

This is a reduction of $100 \times (1.86 - 1.03)/1.86 = 44.7\%$.

If in (b) we assume the sample is from an arbitrary distribution, the width of the confidence interval is

$$2 \times z_{1-\alpha/2}(s/\sqrt{n}) = 2 \times 1.96 \times 2.041/\sqrt{21} = 1.75.$$

Now doubling the sample size changes just the $\sqrt{n}$ part and the width of the confidence interval becomes

$$2 \times z_{1-\alpha/2}(s/\sqrt{n}) = 2 \times 1.96 \times 2.041/\sqrt{42} = 1.23.$$

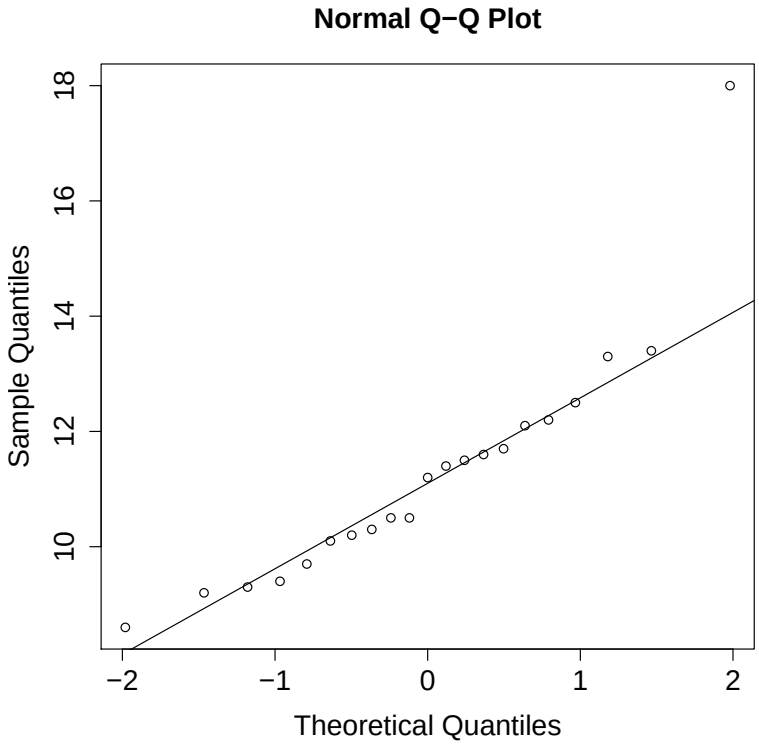
This is a reduction of $100 \times (1.75 - 1.23)/1.75 = 29.3\%$.

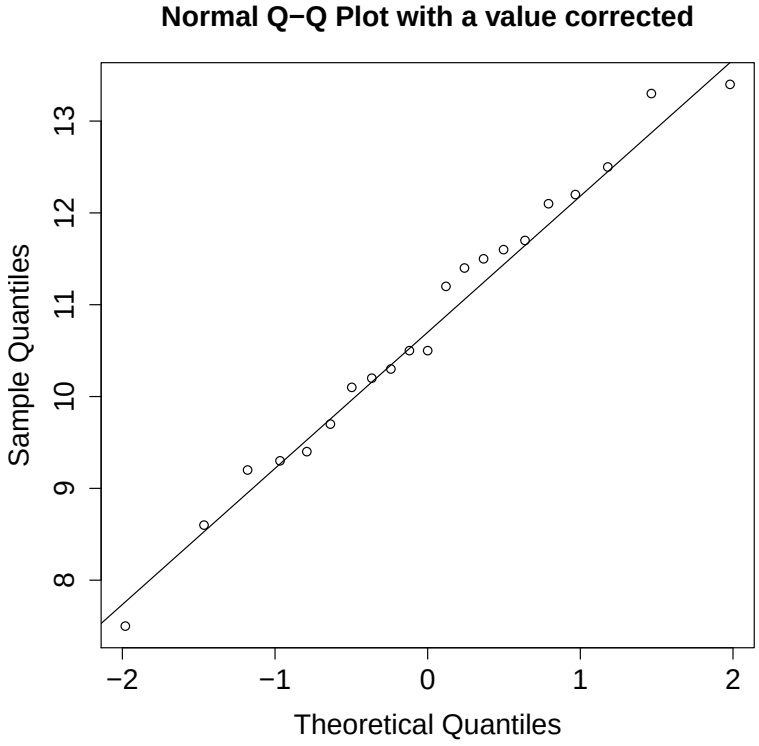
Tripling the sample size the width is $2 \times 1.96 \times 2.041/\sqrt{63} = 1.01$, which is a reduction of $100 \times (1.75 - 1.01)/1.75 = 42.3\%$.

In summary, although the term $t_{n-1,0.975}$ decreases somewhat with increasing sample size, the change in it is relatively small. Most of the change is because of the $1/\sqrt{n}$ part and even then the change is at the rate of $\sqrt{n}$.

(d) The R code I used is given on a subsequent page. (R does have a package, called "boot", for doing boostrapping. But I wanted to demonstrate the principles involved rather than just getting a confidence interval.)

For the particular value in set.seed this yielded 95% CI (10.48, 12.49). Different seeds will yield somewhat different confidence intervals. I tried several values and all gave intervals of the form with lower end 10.4X or 10.5X and upper end 12.4X).


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
10
12
14
16
18
−2
−1
0
1
2


Normal Q−Q Plot with a value corrected
Sample Quantiles
Theoretical Quantiles
8
9
10
11
12
13
−2
−1
0
1
2

R code for the bootstrap-t interval for 1(c):

```
# bootstrap-t interval
set.seed(64353)
mean.Ca<-mean(pge$Ca)
var.Ca<-var(pge$Ca)
n.Ca<-length(pge$Ca)
boots <- 500
zs <- matrix(0,1,boots)
for (jj in 1:boots){
   ysamp <- sample(pge$Ca,size=n.Ca,replace=T)
   zs[jj] <- (mean(ysamp)-mean.Ca)/sqrt(var(ysamp)/n.Ca)
   }
lower.t <- quantile(zs,.975)
upper.t <- quantile(zs,.025)

lower.y <- mean.Ca - lower.t*sqrt(var.Ca/n.Ca)
upper.y <- mean.Ca - upper.t*sqrt(var.Ca/n.Ca)

lower.y
upper.y
```

Programming bootstrap-t intervals in SAS is somewhat more complicated than in R. Below is SAS code, assuming that the data have been read in to a dataset called "pge". For the particular seed used in "proc surveyselect", this yielded 95% CI (10.58, 12.46).

```
proc means data=pge noprint;
  var Ca;
  output out=original mean=original_mean std=original_std n=n;

data original;
  set original;
one=1;  *** For later merging with bootstrap samples;
keep one original_mean original_std n;

proc surveyselect data=pge out=pge_samples seed=45921
    rep=500 sampsize=21 method=urs outhits;
*** rep - specifies the number of replicates
*** method=urs - "requests unrestricted random sampling, which is
***          selection with equal probability and with replacement."
*** outhits - when an observation is selected more than once, the
***          output dataset has a separate row for each occurrence,
***          rather than a single row plus a count of the number of
***          occurrences. ;
```

```
proc means data=pge_samples noprint;
   var Ca;
   by Replicate;
   output out=bootout mean=mean stderr=stderr;

data bootout;
   set bootout;
one=1;

data bootout;
   merge original bootout;
   by one;
zb=(mean-original_mean)/stderr;

proc univariate noprint;
   var zb;
   output out=outpctl pctlpre=P_ pctlpts= 2.5 97.5;

data outpctl;
   set outpctl;
one=1;

data outpctl;
   merge original outpctl;
   by one;

lower=original_mean - P_97_5*original_std/sqrt(n);
upper=original_mean - P_2_5*original_std/sqrt(n);

proc print data=outpctl;
  var lower upper;
```

(e) One way is to do the calculations "manually", using the method on page 43 of the notes on "Point and Interval Estimation". There are 21 patients in the dataset. So we need to find the largest r such that

$$\frac{1}{2^{21}} \sum_{i=0}^{r-1} \binom{21}{i} \leq \alpha/2$$

We can use R to find the largest k such that `sum(dbinom(0:k,21,0.5))` ≤ 0.025 or, equivalently, such that `pbinom(k,21,0.5)` ≤ 0.025.

Because `sum(dbinom(0:5,21,0.5)) = 0.0133` and `sum(dbinom(0:6,21,0.5)) = 0.0392`, $k = 5$ and thus $r - 1 = k = 5$. Hence $r = 6$ and $n - r + 1 = 21 - 6 + 1 = 16$.

A 95% CI for the median is $(X_{(6)}, X_{(16)})$. Sorting the serum calcium values, the 6th and 16th order statistics are 10.1 and 12.1, so the CI is $(10.1, 12.1)$.

Another way is to use SAS:

```
proc univariate cipctldf(type=symmetric);
   var ca;
```

Quantiles (Definition 5)

		95% Conf. Limits		------Order Statistics-----		
Quantile	Estimate	Distribution Free		LCL Rank	UCL Rank	Coverage
100% Max	18.0					
99%	18.0	.	.	.	.	.
95%	13.4	13.3	18.0	19	21	57.45
90%	13.3	12.2	18.0	17	21	83.84
75% Q3	12.1	11.4	13.4	12	20	96.03
50% Median	11.2	10.1	12.1	6	16	97.34
25% Q1	10.1	9.2	10.5	2	10	96.03
10%	9.3	8.6	9.7	1	5	83.84
5%	9.2	8.6	9.3	1	3	57.45
1%	8.6	.	.	.	.	.
0% Min	8.6					

This also has the 95% CI for the median as $(10.1, 12.1)$. The 97.34 in the "Coverage" column is obtained as $100 \cdot (1 - 2 \times 0.0133) = 97.34$.

Question 2 – Problem 4.20 on page 111

(a) If $Y_1, \ldots, Y_n$ is a random sample from a normal distribution with mean μ and variance σ^2, then $\bar{Y} \sim N(\mu, \sigma^2/n)$.

Here $\mu = 1.0$, $\sigma^2 = 9.0$ and $n = 9$, so $\bar{Y} \sim N(1, 9/9) = N(1,1)$. That is, the sampling distribution of $\bar{Y}$ is normal with mean 1 and variance 1.

(b) Standardizing by subtracting the mean of $\bar{Y}$ and dividing by the standard error,

$$\begin{aligned} \Pr[1 < \bar{Y} \leq 2.85] &= \Pr\left[\frac{1-1}{\sqrt{1}} < \frac{\bar{Y}-1}{\sqrt{1}} \leq \frac{2.85-1}{\sqrt{1}}\right] \\ &= \Pr[0 < Z \leq 1.85] = \Pr[Z \leq 1.85] - \Pr[Z \leq 1] \\ &= \Phi(1.85) - \Phi(1) = 0.9678 - 0.5 = 0.4678. \end{aligned}$$

(c) Using properties on page 22 of the notes on "Statistical Inference: Populations and Samples", if $\bar{Y} \sim N(1,1)$ and $W = 4\bar{Y}$, then $W \sim N(4 \times 1, 4^2 \times 1) = N(4, 16)$.

BIOS 662, Fall 2018

Homework 4

Assigned: Thursday, September 27

Due: Thursday, October 4

Instructions: For the problems below, confidence intervals and testing procedures should be done "by hand." You may use appropriate software such as R or SAS to estimate means and variances if these are needed. You should also feel free to use the software to verify any results. For problems involving testing, include a definition of the parameters to be tested, the null and alternative hypotheses, the test statistic to be employed and its distribution, the critical region, whether you reject the null, the p-value, and an interpretation of the results in a language suitable for investigators. (Get into the habit of supplying these, not just for this homework.) All tests should be performed at the $\alpha = 0.05$ significance level.

1. This is based on Problem 5.2 on page 142 of the textbook. "In data of Dobson et al. [1976], 36 patients with a confirmed diagnosis of phenylketonuria (PKU) were identified and placed on dietary therapy before reaching 121 days of age. The children were tested for IQ (Stanford-Binet test) between the ages of 4 and 6; subsequently, their normal siblings of closest age were also tested with the Stanford-Binet." The dataset "HW4_PKU.txt" contains data on the 21 pairs *not* listed in Problem 5.2, which is why numbering of pairs in the dataset starts with 16. For parts (a)–(d) assume IQ data are normally distributed.
   (a) State a suitable null and an alternative hypotheses with regard to these data.
   (b) Test the null hypothesis (using $\alpha = 0.05$).
   (c) Give a 95% confidence interval for the true effect of PKU on IQ.
   (d) State your conclusions.
   (e) What are your assumptions?
   (f) Now suppose we cannot assume normality and need to use the sign test. State the hypotheses, conduct the test and state your conclusions.
   (g) Discuss how and why your conclusions in parts (d) and (f) differ.

2. The following data concern the association between sodium chloride (salt) intake and hypertension. Fifteen hypertensive and twelve normotensive subjects were isolated for a week so that their sodium (Na^+) intakes could be measured accurately. The average daily (Na^+) intakes (in milligrams) are listed in the table below. Compare the average daily (Na^+) intake of the hypertensive subjects with that of the normal volunteers using an appropriate statistical test. Include a justification for the statistical test employed.

Hypertensive	Normal
1100	1000
1320	1220
1350	1300
1450	1400
1600	1555
1850	1600
1900	1780
1990	1780
2050	1900
2120	2020
2200	2350
2210	2375
2500	
2610	
2720	

BIOS 662

Homework 4 Solution

October, 2018

Question 1

The data come from pairs of children (a PKU case and his/her normal sibling). Because the children in a pair are siblings, they cannot be regarded as independent. So it is not appropriate to conduct two-sample tests. Instead, we conduct one-sample tests on the difference between the IQ test scores within each sibling pair.

(a) Let Y_i denote the IQ of the PKU case minus that of his/her normal sibling in pair i. If IQ is normally distributed, then the IQ of the differences is also normally distributed, so it may be reasonable to assume that $Y_1, \ldots, Y_n$ are iid with $Y_i \sim N(\mu, \sigma^2)$, for some μ and σ^2, where μ is the mean difference in IQ between a PKU case and his/her closest-age normal sibling.

Hypotheses: $H_0 : \mu = 0$ versus $H_A : \mu \neq 0$. (A one-sided alternative may also be reasonable, if we think there is no chance the dietary therapy could be so effective as to reverse the direction of association.)

(b) Assuming that the Y_i are normally distributed but that σ^2 is unknown, we use a one-sample t-test. Here $n = 21$, so

$$C_\alpha = \{t : |t| > t_{n-1,1-\alpha/2}\} = \{t : |t| > t_{20,0.975}\} = \{t : |t| > 2.086\}$$

Now

$$t = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} \sim t_{n-1} = \frac{-6.05 - 0}{11.612/\sqrt{21}} = -2.39.$$

Because $|-2.39| = 2.39 > 2.086$, we reject H_0 and conclude that the dietary therapy does not eliminate the IQ gap between cases and their siblings.

Also $p = 2 \cdot \Pr(t_{20} \leq -2.39) = 0.027$.

Using R:

```
> t.test(iq.diff)

        One Sample t-test

data:  iq.diff
t = -2.3866, df = 20, p-value = 0.027
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -11.3335159  -0.7617221
sample estimates:
mean of x
-6.047619
```

Using SAS:

```
proc ttest;
   var iq_diff;

The TTEST Procedure

Variable:  iq_diff

  N        Mean     Std Dev     Std Err     Minimum     Maximum

 21     -6.0476     11.6124      2.5340    -33.0000     20.0000

    Mean       95% CL Mean        Std Dev      95% CL Std Dev

 -6.0476    -11.3335  -0.7617     11.6124      8.8842  16.7691

    DF    t Value    Pr > |t|

    20      -2.39      0.0270
```

(c) A 95% CI for μ is

$$\bar{Y} \pm t_{n-1,1-\alpha/2} \cdot s/\sqrt{n} = -6.05 \pm 2.086 \cdot 11.612/\sqrt{21} = (-11.33, -0.76).$$

This agrees with the results in R and SAS.

(d) Even with dietary therapy, on average a child with PKU has significantly lower IQ at age 4-6 than his or her closest-age normal sibling. The mean IQ of children with PKU who are on the dietary therapy in this study is 6.05 points lower than that of their normal siblings.

(e) Assumptions are that IQ data are normally distributed, that the difference between IQs of pairs all come from the same normal distribution (with the same mean and variance), and that the difference in IQ for any pair is independent of that for any other pair.

(f) Let $\zeta_{0.5}$ denote the median of the differences in IQ between children with PKU and their closest-age normal siblings. Then $H_0 : \zeta_{0.5} = 0$ and $H_A : \zeta_{0.5} \neq 0$.

To determine the critical region we need to find the largest $r_{\alpha/2}$ for which

$$\Pr[R \leq r_{\alpha/2} \mid H_0] = \frac{1}{2^n} \sum_{i=0}^{r_{\alpha/2}} \binom{n}{i} \leq \frac{\alpha}{2}$$

Using R

```
> 2*sum(dbinom(0:5,21,0.5))
[1] 0.0266037

> 2*sum(dbinom(0:6,21,0.5))
[1] 0.07835388
```

Confirming using the SIGN.test function:

```
> SIGN.test(iq.diff)

        One-sample Sign-Test

data:  iq.diff
s = 6, p-value = 0.07835
alternative hypothesis: true median is not equal to 0
```

So $r_{\alpha/2} = 5$ and thus $C_{0.05} = \{0,1,2,3,4,5,16,17,18,19,20,21\}$

In this dataset, $r =$ (number of observations > 0) $= 6 \notin C_{0.05}$ so we cannot reject H_0 and we conclude that the data are consistent with the IQ of the PKU cases being similar to that of their normal siblings. Also, $p = 2 \cdot \Pr(r \leq 6) = 0.078 > 0.05$. Using R:

```
> 2*pbinom(6,21,0.5)
[1] 0.07835388
```

(g) In part (d) we rejected the null hypothesis that the mean difference is zero whereas in (f) we did not reject the null hypothesis that the median difference is 0. When the data are approximately normally distributed the t test can be more powerful than the sign test — the gain in power is because of the additional assumption (normality).

Question 2

Here there is no link between any particular hypertensive and normotensive subjects. So the two samples should be independent and thus two-sample tests should be used.

I argue below for the Wilcoxon test. If you make a reasonable argument for the assumptions of the t-test, it is okay to use it. Below is SAS code for the t-test and corresponding edited output. As with the Wilcoxon test, we would not reject H_0, which in this case is that the mean sodium intake is the same in the two groups.

```
proc ttest;
 class group;
 var sodium;

Variable:  sodium

group           N      Mean   Std Dev   Std Err   Minimum   Maximum

Hypertensive   15    1931.3     490.0     126.5    1100.0    2720.0
Normal         12    1690.0     430.0     124.1    1000.0    2375.0
Diff (1-2)            241.3     464.5     179.9

Method           Variances        DF    t Value    Pr > |t|

Pooled           Equal            25       1.34      0.1918
Satterthwaite    Unequal      24.744       1.36      0.1856

              Equality of Variances

Method      Num DF    Den DF    F Value    Pr > F

Folded F        14        11       1.30    0.6720
```

My preference is to use the Wilcoxon rank sum test here for the following reasons. First, based on the histograms in Figure 1, the sodium intakes in the two groups do not appear to be normally distributed. Second, based on the empirical distribution functions and the boxplots in Figure 1, the assumption of a location shift made by the rank sum test does seem to be plausible.

Let group 1 be the Hypertensive subjects and group 2 the Normal subjects, with sample sizes $n_1 = 15$ and $n_2 = 12$, respectively. Denote the corresponding distribution functions by F_1 and F_2. The null and alternative hypotheses are

$$H_0 : F_1(y) = F_2(y) \quad \text{and} \quad H_A : F_1(y+\Delta) = F_2(y)$$

for all y and some constant $\Delta \neq 0$.

Because n_1 and n_2 are both ≥ 12 the large sample approximation version of the test can be used. The test statistic is

$$Z = \frac{W_1 - E(W_1)}{\sqrt{V(W_1)}}$$

where W_1 is the sum of the ranks from the Hypertension group,

$$E(W_1) = \frac{n_1(N+1)}{2}$$

and

$$V(W_1) = \frac{n_1 n_2 (N+1)}{12} - \frac{n_1 n_2}{12N(N-1)} \sum_{i=1}^{q} t_i(t_i-1)(t_i+1)$$

where $N = n_1 + n_2 = 27$, q denotes the number of sets of ties, and t_i denotes the size of the i^{th} set of ties for $i = 1, \ldots, q$. At the $\alpha = 0.05$ level of significance, the critical region is

$$C_{0.05} = \{Z : |Z| > z_{0.975} = 1.96\}.$$

To compute W_1 we first get the ranks for the observed data, assigning the midrank in the case of ties, as in the table on the next page. There are three sets of ties, each with two tied observations, so $t_i = 2$ in each case.

Here $W_1 = 238$.

Also

$$E(W_1) = \frac{15 \times 28}{2} = 210$$

and

$$V(W_1) = \frac{12 \times 15 \times 28}{12} - \frac{12 \times 15}{12 \times 27 \times 26} \cdot \sum_{i=1}^{3} (2 \times 1 \times 3) = 419.62$$

so that $Z = (238 - 210)/\sqrt{419.62} = 1.367$ and hence we do not reject the null. The p-value is $2 * \Phi(-1.367) = 0.172$. Therefore, there is insufficient evidence from these data to suggest that there is a difference in sodium intake between normal and hypertensive individuals.

Verifying the results using R:

```
> wilcox.test(hypertensive,normal,exact=F,correct=F)

        Wilcoxon rank sum test

data:  hypertensive and normal
W = 118, p-value = 0.1717
alternative hypothesis: true location shift is not equal to 0
```

Hypertensive	Rank	Normal	Rank
1100	2	1000	1
1320	5	1220	3
1350	6	1300	4
1450	8	1400	7
1600	10.5	1555	9
1850	14	1600	10.5
1900	15.5	1780	12.5
1990	17	1780	12.5
2050	19	1900	15.5
2120	20	2020	18
2200	21	2350	23
2210	22	2375	24
2500	25		
2610	26		
2720	27		

Using SAS:

```
proc npar1way wilcoxon correct=no;
 class group;
 var sodium;
```

```
         Wilcoxon Scores (Rank Sums) for Variable sodium
                   Classified by Variable group

                     Sum of    Expected      Std Dev         Mean
group           N    Scores   Under H0     Under H0        Score
-----------------------------------------------------------------
Hypertensive   15     238.0      210.0    20.484516    15.866667
Normal         12     140.0      168.0    20.484516    11.666667

                   Average scores were used for ties.


   Wilcoxon Two-Sample Test

Statistic             140.0000

Normal Approximation
Z                      -1.3669
One-Sided Pr <  Z       0.0858
Two-Sided Pr > |Z|      0.1717
```

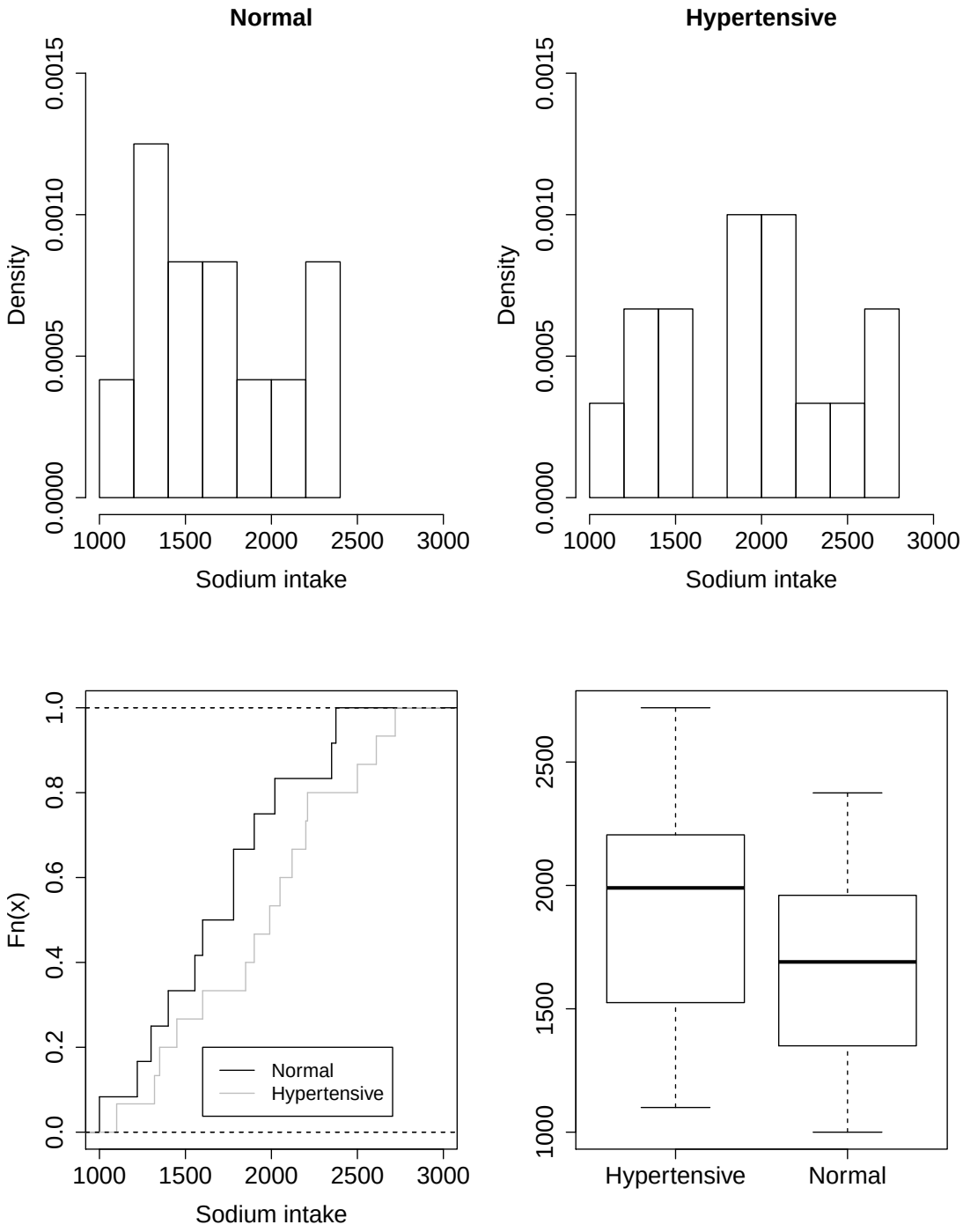


Figure 1: Histograms, EDFs, and boxplots for problem 2

BIOS 662, Fall 2018

Homework 5

Assigned: Tuesday, October 9

Due: Tuesday, October 16

1. This is a continuation of problem #2 from Homework 2 involving the dataset "HW2_PGE.txt". Use the Kolmogorov-Smirnov test to determine whether the distributions of plasma iPGE are the same for people with and without hypercalcemia.

2. Do Problem 6.5 on page 196 of the text. You do not need to do this "by hand" and so can ignore the parenthetical statement about needing to compute hypergeometric probabilities. (In my model solutions I will show how to do it "by hand" though.)

3. Problem 6.11(a)–(c) on page 197 of the text.

4. The file "HW5_Q4.txt" contains the data from Table 6.11 on page 199 of the text in versions suitable for SAS and R (both versions in one text file).
   (a) Verify that collapsing Table 6.11 over the smoking categories yields the table in Problem 6.13.
   (b) Calculate the odds ratio (and 95% confidence interval) for the association between coffee drinking and myocardial infarction, with and without taking into account smoking status. Do the calculations ignoring smoking status "by hand", confirming your results with SAS or R. (The calculations taking smoking status into account do not need to be done "by hand".)
   (c) Does smoking status confound the association between coffee drinking and myocardial infarction?

BIOS 662

Homework 5 Solution

October, 2018

Question 1:

You weren't asked to plot the empirical distribution functions. But it is instructive to see them (and consider ways to plot both in a single graph). The EDFs for the two groups of patients are given in Figure 1. The maximum difference between the two EDFs is indicated by an arrow. One way to obtain EDFs is to use the R function ecdf(...) and then plot the resulting object. To get R to include vertical lines in the plot, in the plot function use the option `verticals=TRUE`. (The default is `verticals=FALSE`.)

For the graph I used the function cumsum to obtain the EDFs "manually" and in the plot function used the option `type="s"` ("stair steps"). Here is my code:

```
ipge_h1<-c(0, 60, 118, 136, 177, 183, 183, 226, 272, 301, 500, 500, 1000)
ipge_h1c<-cumsum(c(0,1,1,1,1,1,1,1,1,1,1,1,0))/11

ipge_h0<-c(0, 88, 100, 121, 130, 144, 148, 150, 168, 172, 254, 1000)
ipge_h0c<-cumsum(c(0,1,1,1,1,1,1,1,1,1,1,1,0))/10

plot(ipge_h1,ipge_h1c,type="s",xlab="Plasma iPGE (pg/mL)",
   ylab="Empirical Distribution Functions F(y)",xlim=c(0,600),lty=2,
   cex.axis=1.25,cex.lab=1.25,cex.main=1.25,cex.sub=1.25)
lines(ipge_h0,ipge_h0c,lty=1,type="s")
legend(350,0.3,c("Hypercalcemia","No hypercalcemia"),lty=c(2,1))
arrows(174.5,0.275,174.5,0.895,col="red",lwd=2,code=3,length=.1)
```

Figure 2 is an alternative version of the plot created using the ecdf(...) function. I haven't been able to find how to suppress the horizontal dashed lines at 0 and 1, which overwrite the parts of the EDFs there.

```
f1=ecdf(c(60, 118, 136, 177, 183, 183, 226, 272, 301, 500, 500))
f2=ecdf(c(88, 100, 121, 130, 144, 148, 150, 168, 172, 254))
plot(f1,verticals=TRUE,pch=NA,ylab="Empirical Distribution Functions F(y)"
  ,xlab="Plasma iPGE (pg/mL)",xlim=c(0,600),lty=2,cex.axis=1.25,
cex.lab=1.25,cex.main=1.25,cex.sub=1.25,ann=FALSE)
lines(f2,lty=1,verticals=TRUE,pch=NA)
legend(350,0.3,c("Hypercalcemia","No hypercalcemia"),lty=c(2,1))
arrows(174.5,0.275,174.5,0.895,col="red",lwd=2,code=3,length=.1)
```

We want to test

$$H_0 : F_1(x) = F_2(x) \text{ for all } x$$

versus

$$H_A : F_1(x) \neq F_2(x) \text{ for at least one } x$$

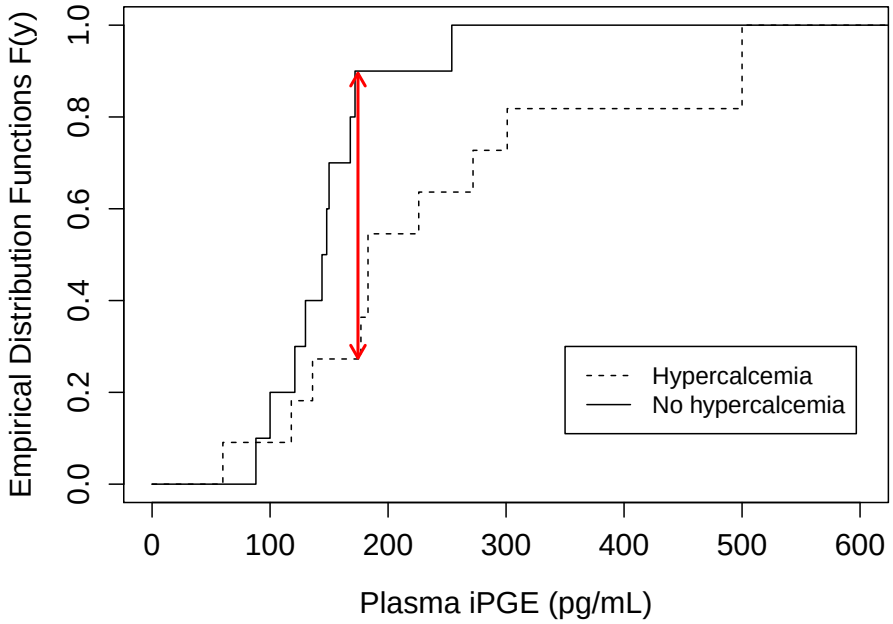


Figure 1: EDFs for problem 1

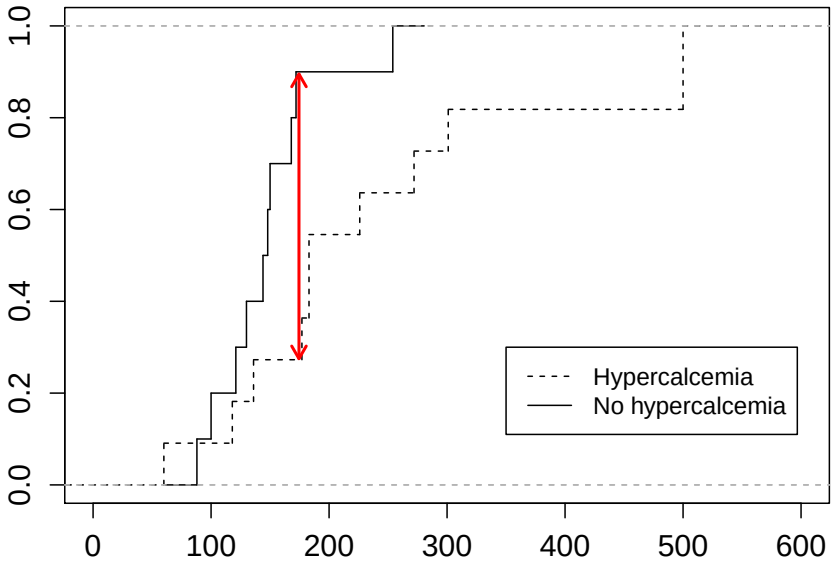


Figure 2: EDFs for problem 1 using ecdf() function

Here $D = \max_x |F_{1n}(x) - F_{2m}(x)| = 9/10 - 3/11 = 0.627$.

From the table on page 268 of the text, $C_{0.05} = \{\text{KS} : \text{KS} \geq 1.36\}$, where KS is defined as

$$\text{KS} = \sqrt{\frac{nm}{n+m}} D = \sqrt{\frac{10 \times 11}{10+11}} \times 0.627 = 1.4356.$$

Thus KS is in $C_{0.05}$ and so we conclude that the distributions of plasma iPGE differ for patients with and without hypercalcemia.

Using SAS to confirm this result and to obtain the p-value (the value for `KSa` is the large-sample approximation):

```
proc npar1way;
   var ipge;
   class hypercalcemia;
   exact ks;

        Kolmogorov-Smirnov Test for Variable iPGE
           Classified by Variable Hypercalcemia

                              EDF at    Deviation from Mean
Hypercalcemia       N       Maximum        at Maximum
------------------------------------------------------------
1                  11      0.272727         -0.990680
0                  10      0.900000          1.039034
Total              21      0.571429

       Maximum Deviation Occurred at Observation 13
             Value of iPGE at Maximum = 172.0

KS  0.3133    KSa  1.4356

Kolmogorov-Smirnov Two-Sample Test

D = max |F1 - F2|      0.6273
Asymptotic Pr >  D     0.0324
Exact      Pr >= D     0.0154
```

Using R:

```
> ks.test(c(60, 118, 136, 177, 183, 183, 226, 272, 301, 500, 500),
+   c(88, 100, 121, 130, 144, 148, 150, 168, 172, 254))

        Two-sample Kolmogorov-Smirnov test

data:  c(60, 118, 136, 177, 183, 183, 226, 272, 301, 500, 500) and
    c(88, 100, 121, 130, 144, 148, 150, 168, 172, 254)
D = 0.6273, p-value = 0.03242
alternative hypothesis: two-sided

Warning message:
In ks.test(c(60, 118, 136, 177, 183, 183, 226, 272, 301, 500, 500),  :
  cannot compute correct p-values with ties
```

Question 2: *Problem 6.5 on page 196 of the text.*

We want to compare the probability of 5-year survival for those with 1–4 courses of chemotherapy to those with ≥ 10 courses. Let $\pi_1 = \Pr[\text{dead}|\text{1–4 courses}]$ and $\pi_2 = \Pr[\text{dead}|\text{10+ courses}]$. Then

$$H_0 : \pi_1 = \pi_2 \text{ and } H_0 : \pi_1 \neq \pi_2.$$

Because the sample size is small we use Fisher's exact test.

To do it "by hand" in the way described in the class notes we first have to rearrange the table so that the row with the smaller row total is the first row and the column with the smaller column total is the first column. That is:

Courses	Alive	Dead	
≥ 10	8	2	10
1–4	2	21	23
Total	10	23	33

Setting $n_{11} = 0$ the table becomes:

Courses	Alive	Dead	
≥ 10	0	10	10
1–4	10	13	23
Total	10	23	33

$$\Pr[n_{11} = 0] = \frac{10!\,23!\,10!\,23!}{33!\,0!\,10!\,10!\,13!} = 0.0124.$$

Next, setting $n_{11} = 1$ the table becomes:

Courses	Alive	Dead	
≥ 10	1	9	10
1–4	9	14	23
Total	10	23	33

$$\Pr[n_{11} = 1] = \frac{10!\,23!\,10!\,23!}{33!\,1!\,9!\,9!\,14!} = 0.0883.$$

Similarly, $\Pr[n_{11} = 2] = 0.2384$, $\Pr[n_{11} = 3] = 0.3178$, $\Pr[n_{11} = 4] = 0.2290$, $\Pr[n_{11} = 5] = 0.0916$, $\Pr[n_{11} = 6] = 0.0201$, $\Pr[n_{11} = 7] = 0.0023$, $\Pr[n_{11} = 8] = 0.0001$, $\Pr[n_{11} = 9] < 0.0001$ and $\Pr[n_{11} = 10] < 0.0001$.

At this point $n_{21} = 0$ and we stop.

a	$\Pr[n_{11} = a]$	$\Pr[n_{11} \leq a]$	$\Pr[n_{11} \geq a]$
0	0.0124	0.0124	1.0000
1	0.0883	0.1006	0.9876
2	0.2384	0.3390	0.8994
3	0.3178	0.6569	0.6610
4	0.2290	0.8859	0.3431
5	0.0916	0.9775	0.1141
6	0.0201	0.9976	0.0225
7	0.0023	0.9999	0.0024
8	0.0001	1.0000	0.0001
9	$<$0.0001	1.0000	$<$0.0001
10	$<$0.0001	1.0000	$<$0.0001

The critical region for $H_A : \pi_1 \neq \pi_2$ is $C_{0.05} = \{n_{11} : n_{11} \in \{0, 6, 7, 8, 9, 10\}\}$. Because $n_{11} = 8$, we reject H_0 and, looking at the observed proportions dying within 5 years ($2/10 = 0.20$ and $21/23 = 0.91$), conclude that survival is more likely among those receiving at least 10 courses of chemotherapy. (Also, $p = 0.0001 < 0.05$.)

We confirm our answer using SAS:

```
data hw5_3;
   input chemo $1-5 status $7-12 count;
datalines;
c10p  alive  8
c10p  dead   2
c1to4 alive  2
c1to4 dead  21
;

proc freq data=hw5_3;
   tables chemo*status / norow nocol nopercent exact;
   weight count;

Chemo     Status

Frequency|alive   |dead    |  Total
---------+--------+--------+
c10p     |      8 |      2 |     10
---------+--------+--------+
c1to4    |      2 |     21 |     23
---------+--------+--------+
Total          10       23       33
```

```
       Fisher's Exact Test
---------------------------------
Cell (1,1) Frequency (F)         8
Left-sided Pr <= F          1.0000
Right-sided Pr >= F      1.255E-04

Table Probability (P)    1.230E-04
Two-sided Pr <= P        1.255E-04
```

Using R:

```
> fisher.test(matrix(c(8,2,2,21),nrow=2))

        Fisher's Exact Test for Count Data

data:  matrix(c(8, 2, 2, 21), nrow = 2)
p-value = 0.0001255
alternative hypothesis: true odds ratio is not equal to 1
```

Question 3: *Problem 6.11(a)-(c) on page 197 of the text.*

From the information given we can set up the table:

Usual church attendance	Arteriosclerotic death: Yes	Arteriosclerotic death: No	
$<$1 per week	89	30,514	30,603
$\geq$1 per week	38	24,207	24,245
Total	127		

Because the hypothesis seems to be that frequent church attendance is associated with "healthier" or "cleaner" living, the more frequent church attendance group is the "unexposed" or lower risk group.

Define $\pi_1 = \Pr[\text{arteriosclerotic death} \mid \text{church } <1 \text{ per week}]$

and $\pi_2 = \Pr[\text{arteriosclerotic death} \mid \text{church } \geq 1 \text{ per week}]$

(a) $\widehat{\text{RR}} = p_1/p_2 = (n_{11}/n_1)/(n_{21}/n_2) = (89/30603)/(38/24245) = 1.8555$

(b) $\widehat{\text{OR}} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}n_{22}}{n_{21}n_{12}} = \frac{89\times 24207}{38\times 30514} = 1.8580$

A 95% CI is $1.8580\exp\left\{\pm 1.96\sqrt{\frac{1}{89}+\frac{1}{38}+\frac{1}{30514}+\frac{1}{24207}}\right\}$

So the confidence interval is (1.270, 2.717).

(c) $100(\widehat{\text{OR}} - \widehat{\text{RR}})/\widehat{\text{RR}} = 100(1.8580 - 1.8555)/1.8555 = 0.13\%$

That is, in this setting in which the disease is rare, the percent error is just a small fraction of a percent.

We confirm parts (a) and (b) using SAS:

```
data;
  input church $1-5 arterio_death $7-9 count;
  datalines;
LT1pw Yes 89
LT1pw No 30514
GE1pw Yes 38
GE1pw No 24207
;

proc freq order=data;
  tables church*arterio_death / nopct nocol norow relrisk;
  weight count;
```

```
church     arterio_death

Frequency|Yes     |No      |  Total
---------+--------+--------+
LT1pw    |     89 |  30514 |  30603
---------+--------+--------+
GE1pw    |     38 |  24207 |  24245
---------+--------+--------+
Total         127    54721    54848

           Estimates of the Relative Risk (Row1/Row2)

Type of Study                   Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)      1.8580        1.2704        2.7174
Cohort (Col1 Risk)             1.8555        1.2696        2.7118
```

Question 4:

(a) *Verify that collapsing Table 6.11 over smoking categories yields the table in Problem 6.13.*

Using SAS on the dataset:

```
proc freq order=data;
   table CupsCoffee*MIcase / norow nocol nopercent;
   weight count;
```

yields the table in Problem 6.13:

```
Table of CupsCoffee by MIcase

CupsCoffee     MIcase

Frequency|Yes     |No      |  Total
---------+--------+--------+
GE5      |    152 |    183 |    335
---------+--------+--------+
LT5      |    335 |    797 |   1132
---------+--------+--------+
Total         487      980     1467
```

(b) *Calculate the odds ratio (and 95% confidence interval) for the association between coffee drinking and myocardial infarction, with and without taking into account smoking status. Do the calculations ignoring smoking status "by hand", confirming your results with SAS or R. (The calculations taking smoking status into account do not need to be done "by hand".)*

Ignoring smoking status, we use the data in the table above.

$\widehat{\text{OR}} = \frac{152\times797}{183\times335} = 1.9761$

A 95% CI is $1.9761 \exp\left\{\pm 1.96\sqrt{\frac{1}{152}+\frac{1}{335}+\frac{1}{183}+\frac{1}{797}}\right\}$

So the confidence interval is (1.5388, 2.5376).

Confirming this using SAS:

```
proc freq order=data;
   table CupsCoffee*MIcase / norow nocol nopercent relrisk;
   weight count;

Statistics for Table of CupsCoffee by MIcase

          Estimates of the Relative Risk (Row1/Row2)

Type of Study                   Value       95% Confidence Limits
-----------------------------------------------------------------
Case-Control (Odds Ratio)      1.9761        1.5388        2.5376
```

Now using the Mantel-Haenszel method to take smoking status into account:

```
proc freq order=data;
   table Smoking*CupsCoffee*MIcase / norow nocol nopercent cmh;
   weight count;

           Estimates of the Common Relative Risk (Row1/Row2)

Type of Study   Method             Value  95% Confidence Limits
---------------------------------------------------------------
Case-Control    Mantel-Haenszel   1.3754    1.0505       1.8007
```

(c) *Does smoking status confound the association between coffee drinking and myocardial infarction?*

There is quite a substantial change in the odds ratio when smoking status is taken into account, decreasing from 1.976 to 1.375. Further evidence of the size of the change is that the latter is below the lower limit of the confidence interval for the former. (Note that this is not a formal test – these are both estimates rather than one being a hypothesized parameter.)

BIOS 662, Fall 2018

Homework 6

Assigned: Tuesday, November 6

Due: Tuesday, November 13

Instructions: For this homework, calculations need not be done "by hand." If you use software such as R or SAS to to the calculations, please include the code you used, not just the output. For all problems involving testing, include a definition of the parameters to be tested, the null and alternative hypotheses, the test statistic to be employed and its distribution, the critical region, whether you reject the null, the p-value, and an interpretation of the results in a language suitable for investigators. All tests should be performed at the $\alpha = 0.05$ significance level.

Peppermint extract is believed to have various medicinal properties. A study was conducted to investigate the effect of peppermint extract on triglyceride levels in rats. Fifty rats were randomly assigned to 5 groups, each having 10 rats. Those in groups 1 through 5 received, respectively, 0 (control group), 75, 150, 300 or 600 mg/kg of peppermint extract daily for three weeks. At the end of three weeks blood was drawn and assayed for various lipids, including triglyerides. The file "HW6_TRG.txt" in the "Datasets" sub-folder of the "Homework materials" section of the Sakai site contains data on triglyceride levels in the blood of the rats. Only those rats with non-missing triglyceride levels are included. There are three variables, ID, group and trg (triglycerides, in μg/dL).

1. The investigators' primary interest is in which groups (that is, which dosages of peppermint extract) differ from one another in terms of the effect on triglyceride levels. Conduct an appropriate statistical analysis of the data using a parametric ANOVA model. Include in your report: (a) an analysis plan, (b) results of your analysis, and (c) a brief conclusion in language suitable for the investigators. As part of your analysis, investigate whether a transformation of the data would be appropriate. If so, state what transformation should be used and check whether it improves the diagnostics, but conduct your analysis on the untransformed data (so as not to introduce an extra level of complication in the grading of this homework).
2. Two items of secondary interest are (i) whether the mean triglyceride level in the control group differs from that in the other 4 groups combined and (ii) whether there is a linear relationship between group number and triglyceride levels. Use your parametric ANOVA model to address these items.
3. Now use a linear regression model with peppermint extract dose as a continuous variable (actual dose in mg/dL, not group number). Provide an

estimate and associated confidence interval for how triglyceride levels change with dose of peppermint extract. Use the regression model to predict the mean triglyceride level for the control group. How does this compare with the sample mean for the control group? (For the purposes of this homework, it is *not* necessary to check the assumptions of the regression model.)

BIOS 662

Homework 6 Solution

November, 2018

Part 1

(a) Analysis Plan

Standard analysis of variance methodology will be used. Triglyceride level Y will be modeled by

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where the index $i = 1,2,3,4,5$ denotes the groups, receiving 0, 75, 150, 300, or 600 mg/kg of peppermint extract, respectively, the index j denotes the j^{th} rat within the i^{th} dosage group, μ_i denotes the population mean triglyceride level in the i^{th} dosage group, and the ε_{ij} are assumed to be independent and identically distributed as $N(0,\sigma^2)$. The primary interest is in pairwise comparisons between groups, to determine which groups differ from one another. The group sizes are unequal, varying from 7 to 10, so it is more appropriate to use the Scheffé or Bonferroni method to adjust for multiple comparisons. We'll use Scheffé's method for the primary analysis.

Standard ANOVA diagnostics will be used to assess the fit of the model above. In the event of violations of the assumptions of ANOVA (in particular, homogeneity of variance or normality), the Box-Cox family of transformations will be used to find the transformation of the data that minimizes the MSE. We recognize that the sample sizes are rather small so that only quite large departures from the assumptions are likely to be detectable.

(b) Analyses

Figure 1 has boxplots of the data for each group, with the individual points overlaid and Table 1 has corresponding summary statistics.

Group	N	Median	Mean	Std Dev
1	10	251.5	244.2	17.87
2	10	241.0	238.1	10.30
3	7	230.0	228.1	8.55
4	10	220.0	220.5	6.67
5	9	210.0	209.9	5.09

Table 1: Summary statistics for data from peppermint extract study

The boxplots, summary statistics and diagnostics for the ANOVA model suggest that the homogeneity of variance assumption is questionable. In particular, looking at the

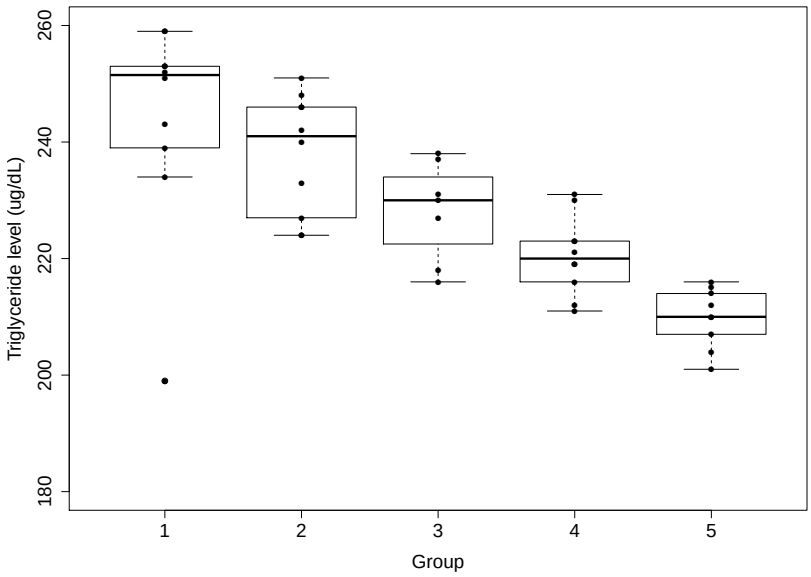


Figure 1: Triglyceride level by peppermint extract dosage group

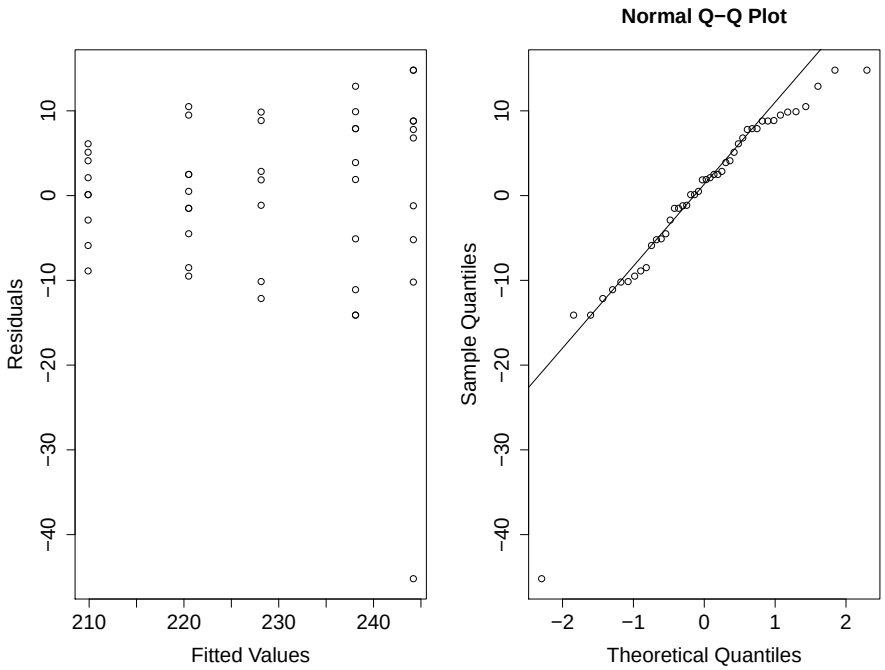


Figure 2: Residual plots from fitted ANOVA model

standard deviations in Table 1 and the residual plot (left panel of Figure 2) it appears that the variance decreases with increasing peppermint extract dose. The trend in the standard deviations is influenced substantially by a single observation, the 199 for rat R15. Omitting this observation reduces the standard deviation for group 1 from 17.86 to 8.70 (and increases that group's mean from 244.2 to 249.2). However, even without excluding this observation, the modified Levene test ($p = 0.38$, see below) does not reject the hypothesis of homogeneity of variance.

```
proc anova;
   class group;
   model trg = group;
   means group / hovtest=bf;

  Brown and Forsythe's Test for Homogeneity of trg Variance
       ANOVA of Absolute Deviations from Group Medians

                       Sum of        Mean
Source        DF     Squares      Square    F Value    Pr > F

group          4       307.6     76.9122       1.09    0.3759
Error         41      2902.5     70.7928
```

In the right panel of Figure 2, the QQ-plot for the residuals from the fitted model suggests that there is departure from normality in the tails of the distribution of the residuals. Pearson's correlation coefficient from the QQ-plot is 0.93 (see below). Here $n = 46$ and as $r = 0.93$ is smaller than the critical value of 0.97 for $n = 40$ on page 9 of the "ANOVA, Part III" overheads, the normality assumption is questionable.

```
> group<-as.factor(group)
> fit<-aov(trg~group)
> par(mfcol=c(1,2))
> plot(fit$fitted.values,fit$residuals)
> qq<-qqnorm(fit$residuals)
> qqline(fit$residuals)
> cor.test(qq$x,qq$y)

        Pearson's product-moment correlation

data:  qq$x and qq$y

sample estimates:
      cor
0.9255097
```

We are not given information about whether some of the rats were from the same litter so we don't have enough information to determine if the independence assumption is satisfied.

Because it appears that the normality assumption is violated and the sample size in each group is rather small for the Central Limit Theorem to help, we will use the Box-Cox

method to investigate potential transformations. From the SAS code and output below, $\lambda = 0.6$ minimizes the MSE. The 95% confidence interval for λ extends from -3.0 to 4.2, indicating a large amount of uncertainty about the most appropriate value for λ. Because a square root transformation ($\lambda = 0.5$) is close to the optimal value, we'll try this to see whether it improves the normality of the residuals. Using this transformation, both parts of Figure 3 are very similar to those in Figure 2, Pearson's correlation coefficient from the QQ-plot is 0.92, little changed from the 0.93 using the untransformed data and the modified Levene test ($p = 0.44$) again does not indicate lack of homogeneity of variance.

(The results of these diagnostic checks are somewhat surprising. For this example I did not have the original data. I saw summary statistics in a journal article and generated data to yield similar summary statistics. I generated the residuals from the normal distribution, with different variances for the 5 groups, yet the diagnostics indicate that normality is questionable rather than homogeneity of variance. So, violation of one of the assumptions may manifest as violation of one of the other assumptions.)

```
data hw6;
   set hw6;

grp1=0; grp2=0;grp3=0;grp4=0;grp5=0;
if group=1 then grp1=1;
   else if group=2 then grp2=1;
   else if group=3 then grp3=1;
   else if group=4 then grp4=1;
   else if group=5 then grp5=1;


%boxcox(resp=trg,model=grp2 grp3 grp4 grp5,lopower=-2,hipower=2,
           npower=41,data=hw6);
```

Box-Cox Power (lambda)	Log Likelihood	Root mean squared error	0.95 Confidence Interval
0.0	-109.931	10.9113	*+
0.1	-109.916	10.9079	*
0.2	-109.905	10.9051	*
0.3	-109.896	10.9030	*
0.4	-109.890	10.9015	*
0.5	-109.886	10.9007	*+
0.6	-109.885	10.9005	<
0.7	-109.887	10.9010	*
0.8	-109.892	10.9021	*
0.9	-109.899	10.9038	*
1.0	-109.909	10.9062	*+

(The output above has been edited to show just part of the range of values of λ.)

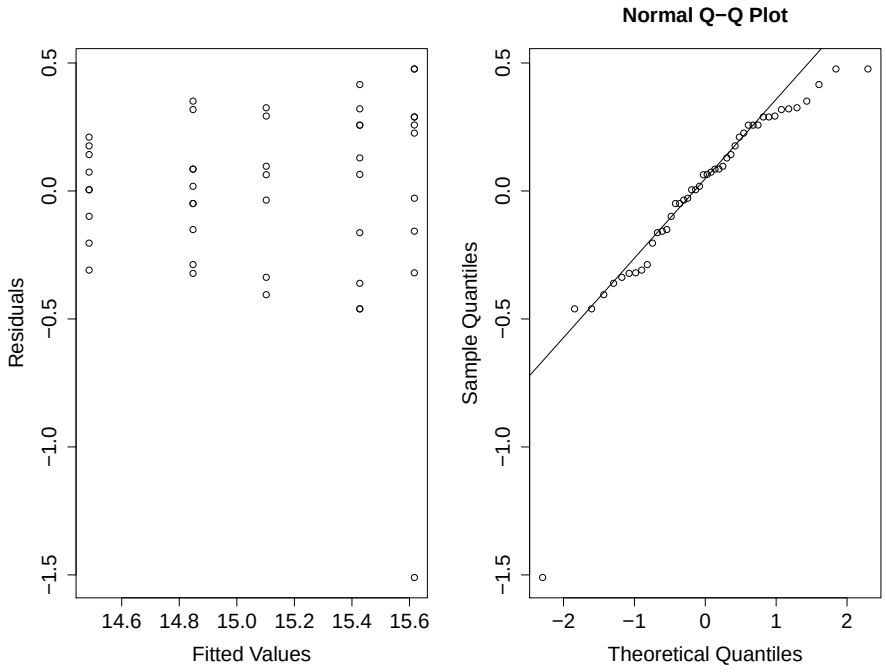


Figure 3: Residual plots from fitted ANOVA model after square root transformation

We now investigate pairwise differences in the means, using the untransformed data, with Scheffé's method for adjusting for the multiple comparisons. SAS code and output are presented below. Groups 1 and 2 differ significantly from groups 4 and 5 and group 3 differs significantly from group 5 but not from the other three groups. These comparisons are presented schematically in Figure 4.

Results using Bonferroni or Tukey's method are similar except that with those methods groups 1 and 3 are significantly different. The corresponding schematic is in Figure 5.

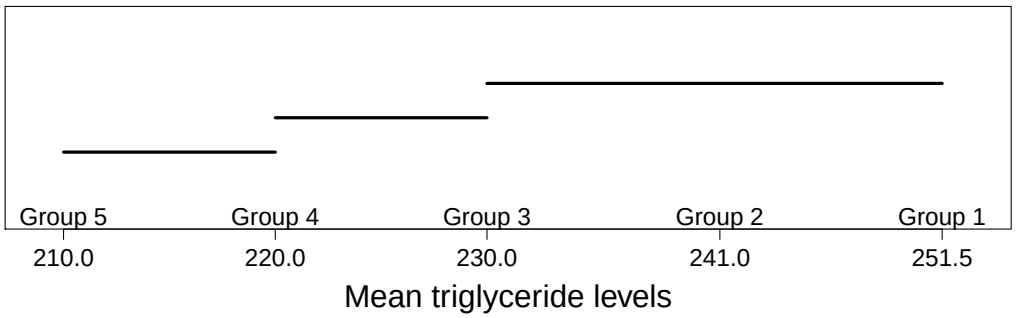


Figure 4: Using Scheffé's method; lines join groups that do not differ significantly

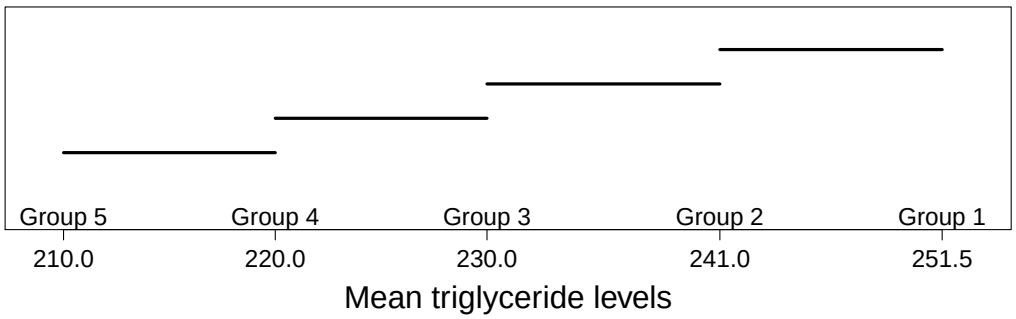


Figure 5: Using Bonferroni; lines join groups that do not differ significantly

```
proc anova data=hw6;
   class group;
   model trg = group;
   means group / scheffe;
```

Dependent Variable: trg

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	7144.55832	1786.13958	15.02	<.0001
Error	41	4876.74603	118.94503		
Corrected Total	45	12021.30435			

Scheffe's Test for trg

NOTE: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than Tukey's for all pairwise comparisons.

Alpha	0.05
Error Degrees of Freedom	41
Error Mean Square	118.945
Critical Value of F	2.59997

Comparisons significant at the 0.05 level are indicated by ***.

group Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
1 - 2	6.100	-9.629	21.829	
1 - 3	16.057	-1.275	33.390	
1 - 4	23.700	7.971	39.429	***
1 - 5	34.311	18.151	50.471	***
2 - 1	-6.100	-21.829	9.629	
2 - 3	9.957	-7.375	27.290	
2 - 4	17.600	1.871	33.329	***
2 - 5	28.211	12.051	44.371	***
3 - 1	-16.057	-33.390	1.275	
3 - 2	-9.957	-27.290	7.375	
3 - 4	7.643	-9.690	24.975	
3 - 5	18.254	0.529	35.979	***
4 - 1	-23.700	-39.429	-7.971	***
4 - 2	-17.600	-33.329	-1.871	***
4 - 3	-7.643	-24.975	9.690	
4 - 5	10.611	-5.549	26.771	
5 - 1	-34.311	-50.471	-18.151	***
5 - 2	-28.211	-44.371	-12.051	***
5 - 3	-18.254	-35.979	-0.529	***
5 - 4	-10.611	-26.771	5.549	

(c) Conclusions

Mean triglyceride levels appear to be included by dosage of peppermint extract, with higher doses associated with lower triglyceride levels. In comparing pairs of dosages, adjacent dosages generally do not have significantly different effects, though that may be because with the small sample sizes there is limited power to detect relatively small differences. Pairs of doages that are further apart generally do result in significant differences in triglyceride levels.

Part 2

Note that the problem specifically says to use your parametric ANOVA model to address these. This can be done using contrasts. For (i) we want to test

$$H_0 : \mu_1 - (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4 = 0 \quad \text{vs.} \quad H_A : \mu_1 \neq (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4$$

or, equivalently,

$$H_0 : \mu_1 = (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4 \quad \text{vs.} \quad H_A : \mu_1 \neq (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4$$

and one way to approach (ii) is to test

$$H_0 : -2 \cdot \mu_1 - 1 \cdot \mu_2 + 0 \cdot \mu_3 + 1 \cdot \mu_4 + 2 \cdot \mu_5 = 0$$

vs.

$$H_A : -2 \cdot \mu_1 - 1 \cdot \mu_2 + 0 \cdot \mu_3 + 1 \cdot \mu_4 + 2 \cdot \mu_5 \neq 0$$

In each case H_0 is of the form $\sum_{i=1}^{5} c_i \mu_i = 0$, with $\sum_{i=1}^{5} c_i = 0$.

Below is SAS code and corresponding output using contrasts to test these hypotheses. In both instances we reject H_0 and conclude that (i) the mean triclyceride level in rats on placebo differs significantly from that in rats given peppermint extract and (ii) there appears to be a linear relationship between group number and mean triclyceride level.

```
proc glm;
   class group;
   model trg = group;
   contrast 'Placebo vs. others' group 1 -0.25 -0.25 -0.25 -0.25;
   contrast 'Linear association' group -2 -1 0 1 2;
```

```
Dependent Variable: trg
                                 Sum of
Source                    DF        Squares    Mean Square    F Value    Pr > F
Model                      4     7144.55832     1786.13958      15.02    <.0001
Error                     41     4876.74603      118.94503
Corrected Total           45    12021.30435

Contrast                  DF    Contrast SS    Mean Square    F Value    Pr > F
Placebo vs. others         1    3129.040058    3129.040058      26.31    <.0001
Linear association         1    7117.919622    7117.919622      59.84    <.0001
```

How does using a contrast in an ANOVA model to test for a linear association differ from using a linear regression model? Below is SAS code and corresponding output from a regression model using group as the predictor variable. The F value is similar but not identical to that for the "Linear association" contrast. The difference arises because the ANOVA model fits 5 parameters (the 5 group means), leaving 41 degrees of freedom for the error whereas the regression model fits two parameters (intercept and slope), leaving 44 degrees of freedom for the error.

```
proc glm;
   model trg = group;

Dependent Variable: trg
                              Sum of
Source              DF       Squares    Mean Square    F Value    Pr > F
Model                1    7094.02783     7094.02783      63.35    <.0001
Error               44    4927.27652      111.98356
Corrected Total     45   12021.30435
```

Part 3

Here we fit a regression model using actual dose rather than group number. The regression model is

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

where X_i is the dose received by the i^{th} rat, Y_i is the triglyceride level of the i^{th} rat and the ε_i are iid $N(0, \sigma^2)$. SAS code and corresponding output are below. Dose is significantly associated with triglyceride levels. The estimated regression model is

$$\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta} X_i.$$

From the output below, $\widehat{\beta} = -0.057$. That is, according to the model, mean triglyceride level decreases by 0.057 μg/dL for each 1 mg/kg increase in peppermint extract dose (or decreases by 5.7 μg/dL for each 100 mg/kg increase in peppermint extract dose). The corresponding 95% confidence interval is

$$\widehat{\beta} \pm t_{N-2,0.975}\text{SE}(\widehat{\beta}) = -0.057 \pm 2.02 \times 0.0076 = (-0.072, -0.041).$$

The predicted mean for the group with dose $= 0$ is $241.1 - 0.0566 \cdot 0 = 241$. Using SAS, the 95% confidence interval for the point on the line at dose $= 0$ is $(236.4, 245.8)$. This compares favorably with the actual mean for group 1, namely 244.2. On the other hand, omitting the outlier rat (R15) the estimate of the mean for dose $= 0$ becomes 243.1, with 95% confidence interval $(239.2, 247.0)$ whereas the mean for group 1 becomes 249.2.

(I hadn't stated explicitly whether the regression should have been done with the observations in the dose $= 0$ group included or removed. The above results are when those observations are included. Fitting the model without the dose $= 0$ observations the predicted mean for dose $= 0$ is 238.6 and the 95% confidence interval for the point on the line at dose $= 0$ is $(233.8, 243.3)$.)

```
proc glm;
   model trg = dose / solution clparm;

Dependent Variable: trg
                                 Sum of
Source                DF        Squares    Mean Square    F Value    Pr > F
Model                  1     6694.23245     6694.23245      55.29    <.0001
Error                 44     5327.07190      121.06982
Corrected Total       45    12021.30435

R-Square     Coeff Var      Root MSE      trg Mean
0.556864      4.814019      11.00317      228.5652

                            Standard
Parameter     Estimate        Error  t Value  Pr > |t|    95% Confidence Limits
Intercept  241.1084873   2.34039305   103.02    <.0001   236.3917350  245.8252395
dose        -0.0565677   0.00760740    -7.44    <.0001    -0.0718994   -0.0412360
```

BIOS 662, Fall 2018

Homework 7

Assigned: Tuesday, November 13

Due: Tuesday, November 20

1. A case-control study is being designed to detect an odds ratio of 3 for bladder cancer associated with a certain medication that is used in about one person out of 50 in the general population. Suppose $\alpha = 0.05$ and that 100 cases and 100 controls are to be sampled for the study. What is the power to detect OR = 3? Would you recommend conducting such a study? If not, how many cases and controls would you recommend?

2. A cross-sectional study is being designed to investigate the association between a continuous exposure X and a continuous outcome Y. The data will be analyzed using a linear regression model

$$Y = \alpha + \beta X + \epsilon.$$

   From a pilot study it seems reasonable to assume that $\epsilon \sim N(0, 10^2)$ and that $X \sim N(50, 8^2)$.

   (a) Using simulation, determine the sample size needed to detect $\beta = 0.20$ with power 0.90. Include a copy of the code you use for your simulation.

   (b) Confirm your answer using a sample size formula from the "Power and Sample Size, Part III" overheads.

BIOS 662

Homework 7 Solution

November, 2018

Question 1

First recast this as a two sample binary problem as in the class notes. Because bladder cancer is rare, the proportion of non-cases in the population is very close to 1. So, to a very good approximation, one person of 50 in the general population using the particular medication is the same as the proportion of controls who use the medication. That is, $\pi_2 = \Pr(\text{exposed}|\text{case}) = 1/50 = 0.02$. Also, as OR $= 3$, we have

$$\pi_1 = \frac{\pi_2 \text{OR}}{1 + \pi_2(\text{OR} - 1)} = 0.05769.$$

The power to detect $\pi_1 = 0.05769$ versus $\pi_2 = 0.02$ at the $\alpha = 0.05$ level of significance is 0.28. This result can be obtained using the formula in the class notes or by the following SAS code:

```
proc power;
   twosamplefreq
     refp   = 0.02
     pdiff  = 0.03769
     ntotal = 200
     power  = .;
```

```
The POWER Procedure
Pearson Chi-square Test for Two Proportions

               Fixed Scenario Elements

Distribution                          Asymptotic normal
Method                              Normal approximation
Reference (Group 1) Proportion                     0.02
Proportion Difference                           0.03769
Total Sample Size                                   200
Number of Sides                                       2
Null Proportion Difference                            0
Alpha                                              0.05
Group 1 Weight                                        1
Group 2 Weight                                        1


Computed Power

Power
0.280
```

We would not recommend conducting this study because the power is very low, so the chance of a type II error is too high. There is a probability of $1 - 0.28 = 0.72$ that we will fail to detect an OR as large as 3. Instead we would recommend 412 cases and 412 controls to have 80% power, or 551 cases and 551 controls to have 90% power. (Or a study with multiple controls per case, but even that would need many more than 100 cases to have reasonable power.) These results can be obtained using the formula in the class notes, as implemented in R in the function "ss_fleiss" defined in the class notes:

```
> ss_fleiss(0.02,0.05769,0.05,0.8)
[1] 411.4046

> ss_fleiss(0.02,0.05769,0.05,0.9)
[1] 550.2548
```

or by using proc power as follows:

```
proc power;
 twosamplefreq
   refp   = 0.02
   pdiff  = 0.03769
   ntotal = .
   power  = 0.8 0.9;
```

```
The POWER Procedure
Pearson Chi-square Test for Two Proportions

               Fixed Scenario Elements

Distribution                         Asymptotic normal
Method                             Normal approximation
Reference (Group 1) Proportion                    0.02
Proportion Difference                          0.03769
Nominal Power                                      0.9
Number of Sides                                      2
Null Proportion Difference                           0
Alpha                                             0.05
Group 1 Weight                                       1
Group 2 Weight                                       1


        Computed N Total

         Nominal    Actual        N
Index      Power     Power    Total
    1        0.8     0.801      824
    2        0.9     0.900     1102
```

Question 2

(a) Using either R or SAS, a key issue for any power/sample size simulation is to work out how to obtain the relevant p-value from the specific function or proc. In the SAS version below the dataset produced by proc reg for the first simulated dataset is printed to check where the p-value is. See the comment in the code. Similarly, in the R code the estimated coefficients and related test statistics are printed for the first simulated dataset.

Note that here for each dataset of size n we generate n values for X and for ε and then obtain n values for Y using $y_i = \alpha + \beta x_i + \varepsilon_i$. In this case any value can be used for α without affecting the results.

Because we want the sample size to achieve a specified power (0.9) we try various values of n until we find one that gives the appropriate power. The appropriate sample size is about 417. (I ran the simulation *before* using the formula in part (b) and will admit to being surprised at how well the simulation agrees with the result in (b).)

For a large sample size and/or large number of simulated datasets, the way the data are generated in the SAS example in the notes can cause out-of-memory errors. So I have included two different approaches to doing the simulation in SAS. The first one is like the example in the notes, with all the datasets being generated first and then the test of the regression coefficient being run on them. In the second approach the datasets are generated one at a time, with the test of the regression coefficient being run before the next dataset is generated. With this approach just one dataset of n observations is kept in memory at any time. The "end=eof", the retain statement and the three lines beginning with "if eof then do;" are to pass an updated random number seed to the next iteration. See pages 20-21 of the "Random Number Generation" overheads.

R code for the simulation and the corresponding output:

```
alpha <- 0
beta <- 0.2

mysim <- function(seed0,n,nsims){
  set.seed(seed0)
  rejects <- 0
  for (ii in 1:nsims){
    e <- rnorm(n,0,10)
    x <- rnorm(n,50,8)
    y <- alpha + beta*x + e
    fit<-summary.lm(lm(y~x))
    coeffs<-fit$coefficients
# The next statement shows the output for the regression on the first simulated
# dataset; looking at the output explains why coeffs[2,4] is used below
    if (ii==1) print(fit$coefficients)
    if (coeffs[2,4]<0.05) rejects <- rejects + 1
    }
  print(paste("Sample size:",n,", estimated power:",rejects/nsims))
  }
```

```
mysim(19,417,10000)
              Estimate Std. Error    t value     Pr(>|t|)
(Intercept) -1.0111480 3.21913633 -0.3141054 0.7535988457
x            0.2191917 0.06362334  3.4451458 0.0006289053
[1] "Sample size: 417 , estimated power: 0.902"
```

Using a different seed yields different estimates for the first dataset of the simulation and a very slightly different power estimate for this sample size:

```
> mysim(37,417,10000)
             Estimate Std. Error   t value   Pr(>|t|)
(Intercept) 2.4377952 3.36466099 0.7245292 0.46914917
x           0.1508311 0.06653812 2.2668369 0.02391407
[1] "Sample size: 417 , estimated power: 0.9005"
```

SAS code using the first approach:

```
%macro epower(beta=,seed=,nsims=,n=);
%let sd_x=8; %let sd_e=10;

data simdata;
  %do i = 1 %to &nsims;
    i=&i;
    do j=1 to &n;
   x=50+rannor(&seed)*&sd_x;
   e=rannor(&seed)*&sd_e;
   y=&beta*x+e;
   output;
end;
  %end;

ods listing close;
ods output "Parameter Estimates"=params;
proc reg data=simdata;
   model y=x;
   by i;
   run;
quit;
ods output close;
ods listing;

*** The following code shows what is in the regression output    ;
*** in "params" dataset when i=1 and thus why "if variable='x'" ;
*** is used below.                                               ;

proc print data=params;
   where i=1;
```

```
data params;
   set params;
if variable='x';

if Probt<0.05 then reject=1; else reject=0;

proc freq data=params;
   table reject;

%mend;

%epower(beta=0.2,seed=97461,nsims=10000,n=417);
```

SAS output from the first approach:

```
Obs i Model  Dependent Variable  DF  Estimate   StdErr  tValue  Probt

  1 1 MODEL1     y     Intercept  1  -1.16503  3.36072   -0.35 0.7290
  2 1 MODEL1     y     x          1   0.22517  0.06601    3.41 0.0007

The FREQ Procedure
                                    Cumulative    Cumulative
reject    Frequency     Percent     Frequency      Percent
-------------------------------------------------------------
     0         998        9.98           998         9.98
     1        9002       90.02         10000       100.00
```

SAS code using the second approach:

```
%let alpha=20; %let beta=0.20; %let sd_e=10;

data pvals; set _NULL_;

%macro rept(seed0=,reps=,sampsize=);
%do i=1 %to &reps;

proc iml;
   setnull=j(&sampsize,1,0);
create begindat from setnull[colname='zero'];
   append from setnull;
quit;

data tempsamp;
   set begindat end=eof;
   retain seed &seed0;

call rannor(seed,z);
call rannor(seed,etemp);
err=&sd_e*etemp;
x=8*z + 50;
y=&alpha + &beta*x + err;
```

```
if eof then do;
call symput('seed0',put(seed,best.));
end;

*** The following code is to check the data being generated ;
*** and look at where the p-value is in the output dataset. ;
%if &i=1 %then %do;

proc means data=tempsamp;
   var z x;

proc reg data=tempsamp outest=temp tableout;
   model y = x;

proc print data=temp;
   var _TYPE_ Intercept x;

%end;

proc reg data=tempsamp outest=regests tableout noprint;
   model y = x;

data regests;
   set regests;
if _TYPE_ NE 'PVALUE' then delete;
p_int=Intercept;
p_x=x;
keep p_int p_x;

data pvals;
   set pvals regests;

%end;
%mend rept;

%rept(seed0=421325,reps=1000,sampsize=417);

data pvals;
  set pvals;
if p_x le 0.05 then reject1=1;
   else reject1=0;

proc freq;
   table reject1;
```

Edited SAS output from the second approach, including PROC MEANS and PROC REG output for the first dataset generated:

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
z	417	-0.0558561	1.0325162	-3.0516687	2.8930635
x	417	49.5531515	8.2601299	25.5866506	73.1445081

The REG Procedure

Number of Observations Read 417
Number of Observations Used 417

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	18.53427	3.10077	5.98	<.0001
x	1	0.22700	0.06173	3.68	0.0003

Obs	_TYPE_	Intercept	x
1	PARMS	18.5343	0.22700
2	STDERR	3.1008	0.06173
3	T	5.9773	3.67756
4	PVALUE	0.0000	0.00027
5	L95B	12.4391	0.10566
6	U95B	24.6295	0.34833

The FREQ Procedure

reject1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	95	9.50	95	9.50
1	905	90.50	1000	100.00

(b) To use the formula from the "Power and Sample Size, Part III" overheads we need s_X and s_Y. Here $s_X = 8$. We are not given s_Y but can calculate it. Because α and β are constants and X and ε are independent:

$$\text{Var}(Y) = \text{Var}(\alpha + \beta \cdot X + \varepsilon) = \beta^2 \text{Var}(X) + \text{Var}(\varepsilon) = 0.2^2 \cdot 8^2 + 10^2 = 102.56.$$

So $s_Y = \sqrt{102.56} = 10.13$.

Letting $\widehat{\beta}_1$ be the alternative we are interested in being able to detect (0.2) we have

$$r = \frac{s_X}{s_Y}\widehat{\beta}_1 = \frac{8}{10.13} \cdot 0.2 = 0.158$$

We should have the same power to test

$$H_A : \beta_1 = 0.2 \quad \text{against} \quad H_0 : \beta_1 = 0$$

as to test

$$H_A : \rho = 0.158 \quad \text{against} \quad H_0 : \rho = 0$$

Here

$$Z_0 = \frac{1}{2}\log\left(\frac{1+0}{1-0}\right) = 0$$

and

$$Z_1 = \frac{1}{2}\log\left(\frac{1+0.158}{1-0.158}\right) = 0.159$$

The sample size n to give us power 0.90 for testing $\rho = 0.159$ versus $\rho = 0$ is

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{(Z_{\rho_1} - Z_{\rho_0})^2} + 3 = \frac{(1.96 + 1.28)^2}{(0.159 - 0)^2} + 3 = 416.9$$

Rounding up, that gives a sample size of $n = 417$.

BIOS 662, Fall 2018

Homework 7

Assigned: Tuesday, November 27

Due: Tuesday, December 4

Calculations need not be done "by hand."

1. This question uses the data in Problem 15.5 on pages 654/5 of the textbook (available on Sakai in the dataset "HW8_Q1.txt", with the "Unknown" category of "Physical Status" coded as 0).

   Note that here the "sample" is the number of operations. These occur at a distinct point in time (and the outcome is vital status at 6 weeks after the operation). Although there is a 4-year time interval, in this particular problem that is essentially irrelevant in terms of incidence. The rate of interest is deaths per 100,000 operations rather than per year (or any other time period).

   SAS versions 9.3 and 9.4 have a procedure for direct and indirect standardization (`PROC STDRATE`). It uses somewhat different formulas for variance estimation from the ones covered in class but you are welcome to use it.

   (a) Calculate the crude death rate (that is, the overall rate, ignoring physical status) per 100,000 operations for halothane and cyclopropane. Are these two rates significantly different?

   (b) Using direct standardization (relative to the total study sample, not just those in the two specific treatment groups), calculate the standardized death rates for halothane and cyclopropane and test whether they are equal.

   (c) Using indirect standardization, with the total study sample as the reference population, calculate the standardized incidence ratio for halothane and test whether $\pi_{halothane}/\pi_{overall} = 1$.

   (Comment: The "total" numbers include those for halothane so the two are not independent, but that complication should be ignored.)

2. The dataset "HW8_SURV.txt" on the Sakai site contains data from a study investigating a new treatment for lung cancer. The variables in the dataset are ID (an identified for each participant), TIMEDEATH (time in days from randomization to death or censoring), DEATH (=0 for a censored observation, =1 for a death), AGE (in years) and GROUP (treatment group;

=1 placebo, =2 the new treatment). The new treatment is intended to be given in addition to usual care. Patients in the placebo group will also be receiving usual care, so the use of a placebo is ethical here.

(a) Compute and plot in the same graph the Kaplan-Meier (product limit) curves for the two treatment groups.

(b) Use the log-rank test to test whether the distribution of survival times is the same in the two groups.

(c) Now use the proportional hazards model to test whether the distribution of survival times is the same in the two groups. That is, use the p-value from the SAS or R output to determine whether the β coefficient differs significantly from 0.

(d) Using your model in part (c), estimate the hazard ratio for group 2 relative to group 1 and provide a 95% confidence interval for the true hazard ratio.

(e) Now include age in the proportional hazards model in part (c). Does age have a significant effect on survival? Does adjusting for age make a substantial difference to the estimate of the treatment effect?

(f) For the placebo group, estimate the median survival time, that is, the time at which $S(t) = 0.5$.

BIOS 662

Homework 8 Solution

December, 2018

Question 1

As mentioned in the question, the "sample" is the number of operations. These occur at a distinct point in time (and the outcome is vital status at 6 weeks after the operation). The rate of interest is deaths per 100,000 operations. We could think of it as "if 100,000 operations were performed in a year, how many people would die within 6 weeks of the operation". Our estimate would be the same if the 100,000 operations occurred over a longer or shorter period than a year. In the example on page 6 of the "Rates and Proportions" overheads, a person without diabetes is at risk of diabetes as long as he/she is being followed in the study. There the incidence of 0.033 can be thought of as the probability of a person becoming diabetic if followed for a year. But in the current problem it is just the short time immediately after the operation that is considered. It does not make sense to try to express this as risk over a year — the risk of death related to the operation decreases with time since the operation, so the risk in the first 6 weeks is unlikely to be representative of the risk over a longer period (such as a year) or even over a shorter period (such as in the first week).

(a) Let $I_{H,10^5}$ and $I_{C,10^5}$ denote the death rates per 100,0000 (per year) for halothane and cyclopropane, respectively.

$$\hat{I}_{H,10^5} = c \cdot \frac{\text{number of deaths}}{\text{number of operations using halothane}}$$

where $c = 100,000$. A corresponding formula holds for $I_{C,10^5}$.

So,

$$\hat{I}_{H,10^5} = 100,000 \cdot \frac{2,375}{146,200} = 1,624.5 \text{ deaths per 100,000 operations per year}$$

and

$$\hat{I}_{H,10^5} = 100,000 \cdot \frac{2,109}{68,169} = 3,093.8 \text{ deaths per 100,000 operations per year.}$$

To test whether the rates differ significantly, that is, $H_0 : I_{H,10^5} = I_{C,10^5}$ versus $H_0 : I_{H,10^5} \neq I_{C,10^5}$, the c terms drop out and we can use a two-sample test of proportions. The sample sizes are large and under H_0,

$$\frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1).$$

Using $\alpha = 0.05$, the critical region is $C_{0.05} = \{|z| > 1.96\}$. Here

$$\frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = \frac{0.0162 - 0.0309}{\sqrt{\frac{0.0162(1-0.0162)}{146200} + \frac{0.0309(1-0.0309)}{68169}}} = -19.8 < -1.96.$$

So we reject H_0 and conclude that the death rate when cyclopropane is used is significantly higher than when halothane is used.

(b) Using all the operations, the weights are given in Table 1, with w_i being the number of operations in physical status category i divided by the total number of operations.

Status	Operations	Weight
Unknown	69,239	0.162
1	185,919	0.435
2	104,286	0.244
3	29,491	0.069
4	3,419	0.008
5	21,797	0.051
6	11,112	0.026
7	2,137	0.005
Total	427,400	1.000

Table 1: Weights for direct standardization

Denote the standardized incidence rates by $I_{H_{adj},10^5}$ and $I_{C_{adj},10^5}$, where for $j \in \{\text{C,H}\}$,

$$\hat{I}_{j_{adj},10^5} = c \cdot \hat{p}_{j_{\text{adj}}} = c \cdot \frac{\sum_{k=1}^{K} w_k \hat{p}_{jk}}{\sum_{k=1}^{K} w_k}.$$

The values of $\hat{p}_{jk}$ are given in Tables 2 and 3 for halothane and cyclopropane, respectively. Using these, $\hat{I}_{H_{adj},10^5} = 100,000 \times 0.018091 = 1,809.1$ deaths per 100,000 operations per year and $\hat{I}_{C_{adj},10^5} = 100,000 \times 0.026843 = 2,684.3$ deaths per 100,000 operations per year.

To test $H_0 : I_{H_{adj},10^5} = I_{C_{adj},10^5}$ versus $H_A : I_{H_{adj},10^5} \neq I_{C_{adj},10^5}$, note that the constant c cancels out again and we have

$$Z = \frac{\hat{I}_{H_{adj},10^5} - \hat{I}_{C_{adj},10^5}}{\sqrt{\widehat{\text{Var}}(\hat{I}_{H_{adj},10^5} - \hat{I}_{H_{adj},10^5})}} \sim N(0,1).$$

Here

$$\widehat{\text{Var}}(\hat{I}_{H_{adj},10^5} - \hat{I}_{H_{adj},10^5}) = \frac{\sum_{k=1}^{K} w_k^2 \left(\widehat{\text{Var}}(\hat{p}_{Hk}) + \widehat{\text{Var}}(\hat{p}_{Ck})\right)}{\left(\sum_{k=1}^{K} w_k\right)^2}$$

with $\widehat{\text{Var}}(\hat{p}_{Hk})$ and $\widehat{\text{Var}}(\hat{p}_{Ck})$ given in Tables 2 and 3 for halothane and cyclopropane, respectively.

So $Z = (0.0181 - 0.0268)/\sqrt{4.95954 \times 10^{-7}} = -12.4 < -1.96$, and as in part (a) we reject H_0.

SAS code for direct standardization plus edited output:

```
proc stdrate data=hw7q2b refdata=hw7q2c method=direct
          effect stat=rate(mult=100000);
    population group=group event=deaths total=ops;
reference total=ops_total;
strata status / stats effect;
```

```
            Directly Standardized Rate Estimates
                   Rate Multiplier = 100000

         --------Study Population-------  -Reference Population-
         Observed  Population-     Crude  Expected   Population-
group      Events          Time      Rate    Events          Time
cyclo        2109         68169    3093.8   11472.6        427400
halo         2375        146200    1624.5    7731.9        427400

     Directly Standardized Rate Estimates
           Rate Multiplier = 100000

         -----------Standardized Rate----------
                    Standard      95% Normal
group    Estimate     Error   Confidence Limits
cyclo      2684.3   62.7938    2561.2    2807.4
halo       1809.1   37.6699    1735.2    1882.9


              Rate Effect Estimates (Rate Multiplier = 100000)

                                      Log
-------group------      Rate         Rate    Standard
   cyclo      halo     Ratio        Ratio       Error          Z    Pr > |Z|

  2684.3    1809.1    1.4838       0.3946      0.0313      12.60      <.0001
```

(c) In calculating the standardized incidence ratio the constant c again cancels out, so we can work with the proportions.

To test $H_0 : \pi_{\text{halothane}}/\pi_{\text{overall}} = 1$ versus $H_A : \pi_{\text{halothane}}/\pi_{\text{overall}} \neq 1$ we first calculate

$$s = \frac{\hat{p}_{\text{halothane}}}{\hat{p}_{\text{overall}}} = \frac{\sum_{k=1}^{K} n_k}{\sum_{k=1}^{K} N_k m_k / M_k} = \frac{O}{E}.$$

If $\widehat{\text{Var}}(O) = \sum_k n_k$ and $\widehat{\text{Var}}(E) = \sum_k \left(\frac{N_k}{M_k}\right)^2 m_k$, then $\widehat{\text{Var}}(s) = \frac{\widehat{\text{Var}}(O) + s^2 \widehat{\text{Var}}(E)}{E^2}$,

Status (k)	Weight (w_k)	Operations	Deaths	$\hat{p}_{Hk}$	$\widehat{\text{Var}}(\hat{p}_{Hk})$
Unknown	0.16200	23684	419	0.01769	0.000000734
1	0.43500	65936	125	0.00190	0.000000029
2	0.24400	36842	560	0.01520	0.000000406
3	0.06900	8918	617	0.06919	0.000007221
4	0.00800	1170	182	0.15556	0.000112272
5	0.05100	6579	74	0.01125	0.000001690
6	0.02600	2632	287	0.10904	0.000036912
7	0.00500	439	111	0.25285	0.000430332
Total	1.00000	146200	2375	—	—

Table 2: Halothane estimates for direct standardization

Status (k)	Weight (w_k)	Operations	Deaths	$\hat{p}_{Ck}$	$\widehat{\text{Var}}(\hat{p}_{Ck})$
Unknown	0.16200	10147	297	0.02927	0.000002800
1	0.43500	27444	91	0.00332	0.000000120
2	0.24400	14097	361	0.02561	0.000001770
3	0.06900	3814	403	0.10566	0.000024777
4	0.00800	681	127	0.18649	0.000222778
5	0.05100	7423	101	0.01361	0.000001808
6	0.02600	3814	476	0.12480	0.000028639
7	0.00500	749	253	0.33778	0.000298646
Total	1.00000	68169	2109	—	—

Table 3: Cyclopropane estimates for direct standardization

and under H_0, $Z = (s-1)/\sqrt{\widehat{\text{Var}}(s)} \sim N(0,1)$.

Using the data in Table 4, $O = 2,375$, $E = 2,695.12$, $s = 2,375/2,695.12 = 0.88$, $\widehat{\text{Var}}(O) = 2,375$ and $\widehat{\text{Var}}(E) = 848.27$.

So $\widehat{\text{Var}}(s) = \dfrac{2375 + 0.88^2 \cdot 848.27}{2695.12^2} = 0.00043$ and $Z = \dfrac{0.88-1}{\sqrt{0.00043}} = -5.73 < -1.96$.

Thus we reject H_0 and we conclude that the mortality rate for halothane is significantly less than the overall death rate.

SAS uses a somewhat different estimator for the variance. It uses

$$\widehat{\text{Var}}(s) = \frac{O}{E^2} = \frac{2375}{2695.12^2} = 0.00033$$

and this yields

$$Z = \frac{0.88-1}{\sqrt{0.00033}} = -6.57.$$

```
proc stdrate data=hw7q2 refdata=hw7q2 method=indirect stat=rate;
    population event=dth_halo total=ops_halo;
reference event=dth_total total=ops_total;
strata status / stats smr;
```

Standardized Morbidity/Mortality Ratio

Observed Events	Expected Events	SMR	Standard Error	95% Normal Confidence Limits		Z	Pr > |Z|
2375	2695.12	0.8812	0.0181	0.8458	0.9167	-6.57	<.0001

	Reference		Halothane			
Status (k)	m_k	M_k	n_k	N_k	$N_k m_k / M_k$	$\left(\frac{N_k}{M_k}\right)^2 m_k$
Unknown	1,378	69,239	419	23,684	471.36	161.23
1	445	185,919	125	65,936	157.82	55.97
2	1,856	104,286	560	36,842	655.68	231.64
3	2,135	29,491	617	8,918	645.62	195.23
4	590	3,419	182	1,170	201.90	69.09
5	314	21,797	74	6,579	94.77	28.61
6	1,392	11,112	287	2,632	329.71	78.10
7	673	2,137	111	439	138.25	28.40
Total	8,783	427,400	2,375	146,200	2695.12	848.27

Table 4: Counts for indirect standardization

Question 2

(a) Tables 5 and 6 give the calculations for the Kaplan-Meier curves for group 1 and group 2, respectively.

Figure 1 has the Kaplan-Meier survival function estimates for the two groups, plotted using the R code:

```
library("survival")

fit <- survfit(Surv(timedeath, death)~group,conf.type="none")

pdf("HW8_Surv.pdf",width=11,height=8.5)

plot(fit,xlab="Time (days)",ylab="S(t)",lwd=c(1,3),cex.axis=1.6,
   main="Kaplan-Meier estimates for the two groups",cex.lab=1.6,
 cex.main=1.6,cex.sub=1.6)

legend(425,1.0,c("New treatment","Placebo"),lwd=c(3,1),cex=1.6)

dev.off()
```

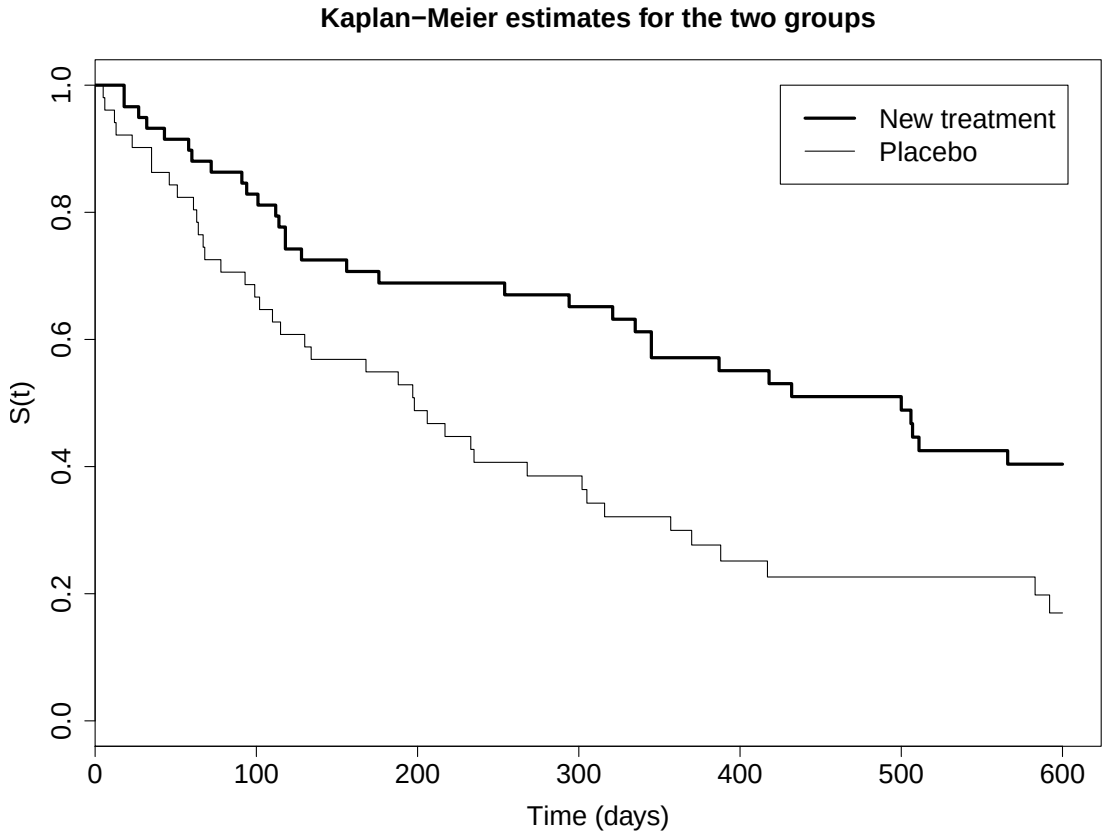


Figure 1: Calculation of Kaplan-Meier estimate for group 1

$t_{(j)}$	m_j	q_j	$R(t_{(j)})$	$\hat{S}(t_{(j)})$
5	1	0	51	0.98039
6	1	0	50	0.96078
12	1	0	49	0.94118
13	1	0	48	0.92157
23	1	0	47	0.90196
35	2	0	46	0.86275
46	1	0	44	0.84314
51	1	0	43	0.82353
61	1	0	42	0.80392
63	1	0	41	0.78431
64	1	0	40	0.76471
67	1	0	39	0.74510
68	1	0	38	0.72549
78	1	0	37	0.70588
93	1	0	36	0.68627
99	1	0	35	0.66667
102	1	0	34	0.64706
110	1	0	33	0.62745
115	1	0	32	0.60784
130	1	0	31	0.58824
134	1	0	30	0.56863
168	1	1	29	0.54902
188	1	0	27	0.52869
197	1	0	26	0.50835
198	1	0	25	0.48802
206	1	0	24	0.46768
217	1	0	23	0.44735
233	1	0	22	0.42702
235	1	1	21	0.40668
268	1	0	19	0.38528
302	1	0	18	0.36387
305	1	0	17	0.34247
316	1	0	16	0.32106
357	1	1	15	0.29966
370	1	1	13	0.27661
388	1	0	11	0.25146
417	1	1	10	0.22632
583	1	0	8	0.19803
592	1	6	7	0.16974

Table 5: Calculation of Kaplan-Meier estimate for group 1

$t_{(j)}$	m_j	q_j	$R(t_{(j)})$	$\hat{S}(t_{(j)})$
18	2	0	59	0.96610
27	1	0	57	0.94915
32	1	0	56	0.93220
43	1	1	54	0.91494
58	1	0	53	0.89768
60	1	0	52	0.88041
72	1	0	51	0.86315
91	1	0	50	0.84589
94	1	0	49	0.82863
101	1	0	48	0.81136
112	1	0	47	0.79410
114	1	0	46	0.77684
118	2	0	45	0.74231
128	1	2	43	0.72505
156	1	0	40	0.70692
176	1	1	39	0.68879
254	1	0	37	0.67018
294	1	2	36	0.65156
321	1	0	33	0.63182
335	1	0	32	0.61207
345	2	0	30	0.57127
387	1	0	28	0.55087
418	1	0	27	0.53046
432	1	1	26	0.51006
500	1	0	24	0.48881
506	1	0	23	0.46756
507	1	0	22	0.44630
511	1	0	21	0.42505
566	1	19	20	0.40380

Table 6: Calculation of Kaplan-Meier estimate for group 2

(b) Tables 7 and 8 have data for doing the log-rank test "by hand". We want to test $H_0 : S_1(t) = S_2(t)$ for all t against $H_A : S_1(t) \neq S_2(t)$ for at least one t. Under H_0, $X = (O_1 - E_1)^2/V_1 \sim \chi^2_1$. So the critical region is $C_{0.05} = \{X : X > \chi^2_{1,0.95} = 3.84\}$.

Using the data in Tables 7 and 8:

$$X = (O_1 - E_1)^2/V_1 = (40 - 27.8195)^2/16.7682 = 8.85 > 3.84.$$

So we reject H_0 and conclude that the new treatment tends to increase survival time in comparison with placebo. We confirm the result using SAS:

```
proc lifetest;
   time timedeath*death(0);
   strata group;

       Test of Equality over Strata
                                    Pr >
Test      Chi-Square      DF    Chi-Square

Log-Rank      8.8468       1      0.0029
```

Note that if one does the calculations using group 2 in X one obtains the same value for the statistic. In that case $O_2 = 32$ and $E_2 = 44.1805$.

(c) Let X be an indicator of being in group 2, that is $X = 0$ if in group 1 and $X = 1$ if in group 2. The model is

$$\log \lambda(t) = \log \lambda_0(t) + \beta X$$

We want to test $H_0 : \beta = 0$ versus $H_A : \beta \neq 0$. We have not covered details of how the test is conducted and so will use just the output from SAS or R. Using SAS with the option "rl" in the "model" statement to obtain the "risk limits", that is, a 95% confidence interval for the hazard ratio, needed in part (d):

```
proc phreg;
   model timedeath*death(0) = group01 / ties=exact rl;


The PHREG Procedure

               Analysis of Maximum Likelihood Estimates

                    Parameter      Standard
Parameter    DF      Estimate         Error    Chi-Square    Pr > ChiSq

group01       1      -0.69719       0.23939        8.4820        0.0036

  Analysis of Maximum Likelihood Estimates

               Hazard      95% Hazard Ratio
Parameter       Ratio      Confidence Limits

group01         0.498       0.311       0.796
```

$t_{(k)}$	m_{1k}	$R_1(t_{(k)})$	m_{2k}	$R_2(t_{(k)})$	m_k	$R(t_{(k)})$	E_{1k}	V_{1k}
5	1	51	0	59	1	110	0.46364	0.24868
6	1	50	0	59	1	109	0.45872	0.24830
12	1	49	0	59	1	108	0.45370	0.24786
13	1	48	0	59	1	107	0.44860	0.24736
18	0	47	2	59	2	106	0.88679	0.48889
23	1	47	0	57	1	104	0.45192	0.24769
27	0	46	1	57	1	103	0.44660	0.24715
32	0	46	1	56	1	102	0.45098	0.24760
35	2	46	0	55	2	101	0.91089	0.49107
43	0	44	1	54	1	98	0.44898	0.24740
46	1	44	0	53	1	97	0.45361	0.24785
51	1	43	0	53	1	96	0.44792	0.24729
58	0	42	1	53	1	95	0.44211	0.24665
60	0	42	1	52	1	94	0.44681	0.24717
61	1	42	0	51	1	93	0.45161	0.24766
63	1	41	0	51	1	92	0.44565	0.24705
64	1	40	0	51	1	91	0.43956	0.24635
67	1	39	0	51	1	90	0.43333	0.24556
68	1	38	0	51	1	89	0.42697	0.24467
72	0	37	1	51	1	88	0.42045	0.24367
78	1	37	0	50	1	87	0.42529	0.24442
91	0	36	1	50	1	86	0.41860	0.24337
93	1	36	0	49	1	85	0.42353	0.24415
94	0	35	1	49	1	84	0.41667	0.24306
99	1	35	0	48	1	83	0.42169	0.24387
101	0	34	1	48	1	82	0.41463	0.24271
102	1	34	0	47	1	81	0.41975	0.24356
110	1	33	0	47	1	80	0.41250	0.24234
112	0	32	1	47	1	79	0.40506	0.24099
114	0	32	1	46	1	78	0.41026	0.24195
115	1	32	0	45	1	77	0.41558	0.24287
118	0	31	2	45	2	76	0.81579	0.47659
128	0	31	1	43	1	74	0.41892	0.24343
130	1	31	0	42	1	73	0.42466	0.24432
134	1	30	0	42	1	72	0.41667	0.24306
156	0	29	1	40	1	69	0.42029	0.24365
168	1	29	0	39	1	68	0.42647	0.24459
176	0	28	1	39	1	67	0.41791	0.24326
188	1	27	0	38	1	65	0.41538	0.24284

Table 7: First part of table for log-rank test

$t_{(k)}$	m_{1k}	$R_1(t_{(k)})$	m_{2k}	$R_2(t_{(k)})$	m_k	$R(t_{(k)})$	E_{1k}	V_{1k}
197	1	26	0	38	1	64	0.40625	0.24121
198	1	25	0	38	1	63	0.39683	0.23936
206	1	24	0	38	1	62	0.38710	0.23725
217	1	23	0	38	1	61	0.37705	0.23488
233	1	22	0	38	1	60	0.36667	0.23222
235	1	21	0	38	1	59	0.35593	0.22924
254	0	20	1	37	1	57	0.35088	0.22776
268	1	19	0	36	1	55	0.34545	0.22612
294	0	18	1	36	1	54	0.33333	0.22222
302	1	18	0	35	1	53	0.33962	0.22428
305	1	17	0	35	1	52	0.32692	0.22004
316	1	16	0	35	1	51	0.31373	0.21530
321	0	15	1	33	1	48	0.31250	0.21484
335	0	15	1	32	1	47	0.31915	0.21729
345	0	15	2	30	2	45	0.66667	0.43434
357	1	15	0	28	1	43	0.34884	0.22715
370	1	13	0	28	1	41	0.31707	0.21654
387	0	12	1	28	1	40	0.30000	0.21000
388	1	11	0	27	1	38	0.28947	0.20568
417	1	10	0	27	1	37	0.27027	0.19722
418	0	9	1	27	1	36	0.25000	0.18750
432	0	9	1	26	1	35	0.25714	0.19102
500	0	9	1	24	1	33	0.27273	0.19835
506	0	9	1	23	1	32	0.28125	0.20215
507	0	9	1	22	1	31	0.29032	0.20604
511	0	9	1	21	1	30	0.30000	0.21000
566	0	9	1	20	1	29	0.31034	0.21403
583	1	8	0	19	1	27	0.29630	0.20850
592	1	7	0	19	1	26	0.26923	0.19675
	40		32				27.8195	16.7682

Table 8: Second part of table for log-rank test

Equivalently, in R:

```
> coxph(Surv(timedeath, death)~group01)
> summary(coxph(Surv(timedeath, death)~group01))

Call:
coxph(formula = Surv(timedeath, death) ~ group)

  n= 110, number of events= 72

         coef exp(coef) se(coef)      z Pr(>|z|)
group -0.6972    0.4980   0.2394 -2.913  0.00358 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

      exp(coef) exp(-coef) lower .95 upper .95
group     0.498      2.008    0.3115    0.7961
```

In either case the p-value associated with the test of $\beta = 0$ is 0.0036. So we reject H_0 and we conclude that the time to death differs significantly between the two treatment groups, with the new treatment being better than placebo.

(d) The hazard ratio estimate is $\exp(\hat{\beta}) = \exp(-0.6972) = 0.498$. The 95% CI can be obtained from SAS or R output. Alternatively, as in Table 16.7 of the text, a 95% CI for β is given by $\hat{\beta} \pm 1.96 \cdot \text{SE}(\hat{\beta}) = -0.6972 \pm 1.96 \cdot 0.2394 = (-1.166, -0.228)$. Taking antilogs gives the 95% CI for the hazard ratio as $(0.311, 0.796)$.

(e) The model is now

$$\log \lambda(t) = \log \lambda_0(t) + \beta_{group} X_{group} + \beta_{age} X_{age}$$

and the SAS code and output as below. The p-value for age is 0.0339, so age has a significant efect on survival. The associated hazard ratio is greater than one, so the hazard increases with age, that is, older age is associated with poorer survival probability. The hazard ratio for the treatment group variable has not changed substantially.

```
proc phreg;
   model timedeath*death(0) = group01 age / ties=exact rl;

The PHREG Procedure

                  Analysis of Maximum Likelihood Estimates

                Parameter   Standard                              Hazard
Parameter  DF    Estimate      Error   Chi-Square   Pr > ChiSq     Ratio

group01     1    -0.72071    0.24017       9.0047       0.0027     0.486
age         1     0.04922    0.02320       4.5012       0.0339     1.050
```

(f) Looking at Table 5, the time at which $\widehat{S}(t)$ is first ≤ 0.5 is at $t = 198$.

BIOS 662, Fall 2018
Midterm Examination

Assigned: Tuesday, October 23

Due: 11:59 PM on Thursday, November 1

Instructions:

The midterm exam is a take-home exam. It is due just before midnight on Thursday, November 1. Please put your completed exam in the BIOS662 mailbox in the Department of Biostatistics. (I will collect the exams from the mailbox *very* early on the Friday.)

Do not discuss the exam with anyone. If you need clarification on any question you may contact me by email (david_couper@unc.edu). The graders will not answer questions about the exam but you may continue to ask them more general questions during their office hours.

Please sign your name on the exam indicating that you did not receive assistance from anyone. Note that obtaining help from anyone other than me is an Honor Code violation.

You may use software to perform calculations but make sure that you include enough information in your answers that I can see what you have done. For instance, include the SAS or R statements you used, not just the output. For all problems involving statistical testing, include a definition of the parameters to be tested, the null and alternative hypotheses, the test statistic to be employed and its distribution (when this is relevant), whether you reject the null, and *an interpretation of the results in a language suitable for investigators.* You may do tests either by determining whether the test statistic falls in the critical region or by obtaining a p-value (you don't have to use both approaches). For any method you use, be sure to state and check the required assumptions. All tests should be performed at the $\alpha = 0.05$ significance level.

Datasets and manuscripts referred to in the questions are available on the Sakai web site, in the folder "Midterm materials" under "Resources". All data are fictitious, even when based on real studies.

1. Write and sign a statement confirming you have not obtained assistance from anyone.

2. The gestational age (GA) of an embryo or newborn infant is approximately the time since the mother's last menstrual period. GA is used to determine whether the pregnancy has reached full term (40 weeks). An infant born prior to 37 weeks GA is regarded as being premature or preterm. There has been some epidemiologic evidence for an association between a pregnant woman having periodontal disease and giving birth prematurely or for the infant's birthweight to be below normal.

   The MOTOR Study was a randomized clinical trial to investigate whether treating periodontal disease in pregnant women reduces the risk of preterm delivery and/or increases average birthweight. (The manuscript by Offenbacher *et al.* assigned in homework 1 provides information about the study. It is not necessary to consult that manuscript in order to complete this exam.)

   In MOTOR about 1,800 pregnant women with periodontal disease were randomized into two treatment groups. The "prenatal treatment" group received periodontal therapy early in pregnancy. The "post-partum treatment" group received periodontal therapy a few weeks after delivery.

   Two measures of GA at birth were available. The more accurate one was made using an ultrasound examination early in the pregnancy. This is given in weeks, calculated from days. So, for instance, 38.1429 corresponds to a GA of 38 weeks and 1 day. The other measure of GA was an estimate made when the infant was born and is in whole weeks.

   Various periodontal measurements were made around each tooth at baseline (early in the pregnancy, before randomization) and were repeated shortly after giving birth. For each woman, the measurements on the teeth were averaged to give a summary score for each type of measurement. One such measurement is probing depth, in millimeters, with larger values indicating more periodontal disease.

   The file "Midterm_BWT.dat" and corresponding SAS and R datasets contain data from a subset of the live births in the trial. The columns in the file are, respectively, ID (participant identifier), group (treatment group; 1 = prenatal, 2 = post-partum), rand_month (month in which the woman was randomized, with 1=January, 2=February, etc.), birth_month (month in which the woman gave birth), GA_ultra (GA estimated by ultrasound), GA_est (GA estimated at birth), ppnum (number of previous pregnancies, as a character variable, with ≥ 3 denoted as "3+"), PD_pre (average pocket depth at the time of randomization), PD_post (average pocket depth after delivery).

   If you detect any apparently incorrect data values, make a note of these, explain why you believe the data are incorrect, and set the values to missing. Regard a

value as being incorrect only if it is clearly impossible. You may assume that there are no errors in the pocket depth variables. *Set incorrect values to missing.*

(a) Is the ultrasound version of GA approximately normally distributed?

(b) Do the means of the two gestational age variables differ?

(c) After taking into account any difference in the means (whether or not statistically significant), do the shapes of the distributions of the two gestational age variables differ?

(d) Classify both versions of gestational age into 3 intervals, $(0, 37)$, $[37, 40)$, and $[40, \infty)$. Determine how well the two versions agree and provide a 95% confidence interval for the true agreement.

(e) Is the number of women randomized in each month consistent with the number of days in each month?

(f) Without doing any additional tests, comment on how the distribution of the number of births each month compares with that of the number of women randomized each month.

For parts (g) and (h), dichotomize the ultrasound version of GA to define preterm delivery.

(g) Does the risk of preterm delivery vary monotonically with the number of previous pregnancies?

(h) Based on this study, is treating periodontal disease in pregnant women effective in terms of reducing the risk of prematurity?

The effect of the periodontal therapy on mean birthweight was smaller than the investigators had expected – there was not a statistically significant difference between the mean birthweights in the two treatment groups. One potential explanation for the lack of effect is that the periodontal therapy provided may not have been intensive enough to yield a substantial and sustained reduction in the amount of periodontal disease.

(i) Ignoring treatment group, is there a difference between the mean average pocket depth at randomization and the mean average pocket depth after delivery.?

(j) Did the mean change in average pocket depth differ between the two treatment groups?

(k) Based on the data on average pocket depth, discuss the effectiveness of the periodontal therapy and the consequences for the potential to affect birthweight.

3. The evidence on which the MOTOR Study was based includes data from case-control studies. The file "Midterm_CC.dat" contains data from one such study. Cases were defined as women who had given birth prematurely (< 37 weeks GA). Controls had full-term babies. The women had a periodontal examination soon after giving birth. Those who had moderate or severe periodontal disease (based on the investigators' criteria) were classified as "exposed" to periodontal disease and those who had no evidence of periodontal disease or just mild disease were classified as "unexposed". Age was considered to be a potential confounder of the association.

   The columns in the file are, respectively, ID (participant identifier), case (indicator of premature birth case status; 1 = case, 0 = control), exposed (indicator of periodontal disease status; 1 = moderate or severe periodontal disease, 0 = no more than mild periodontal disease), and age_group (the age group of the mother, with 1 representing the youngest age group and 3 the oldest). You may assume there are no errors in this dataset.

   First assume this was an unmatched case-control study.

   (a) Determine whether premature birth case status is associated with being exposed to periodontal disease.

   (b) Provide an estimate for a measure of the association between exposure and case status and give a 95% confidence interval for the true measure.

   (c) Repeat part (b) above, taking age group into account.

   (d) Does age group appear to be a confounder? Is the pooled estimate in part (c) a reasonable way to summarize the association here?

   The data were actually from an individually-matched case-control study, with one control per case, matching on age and number of previous pregnancies. The first 4 characters of the ID indicate the case-control pair (the case and the control in the pair have the same first 4 characters). The final character of the ID is 1 for the case in the pair and 0 for the control.

   (e) Repeat parts (a) and (b) above assuming an individually-matched case-control design.

   (f) Which of the estimates of the measure of association in (b), (c) and (e) is most appropriate? Justify your choice.

   (g) Discuss the strength of the evidence from this case-control study for an association between periodontal disease and preterm delivery.

BIOS 662, Fall 2018
Solution to Midterm Exam

Question 1

Question 2

Investigation of potentially incorrect data values: Below is a list of reasons for setting various values to missing, with the IDs of the corresponding women given in parentheses.

- A gestational age of 75 is impossible (GA_ultra for M2349).
- Months are numbered 1 through 12, so rand_month values of 0 (M1190) and 15 (M1722) are errors.
- Number of previous pregnancies cannot be negative, so values of -9 are errors (M1190 and M1410).

(a) A histogram, stem-and-leaf plot or boxplot suggests substantial skewness in the data, with a long tail towards lower values of GA. A QQ plot also shows substantial departure from normality See, for instance, Figure 1. To test for normality, use the Kolmogorov-Smirnov test. We need the Lilliefors version to adjust for having to estimate the mean and variance.

H_0 : GA_est is normally distributed; H_A : GA_est is not normally distributed.

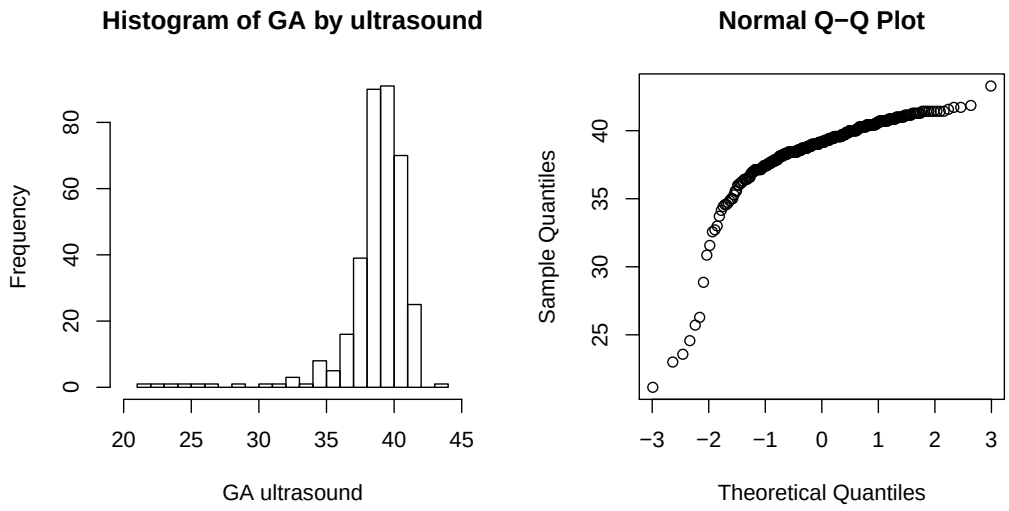
SAS gives the p-value for the Lilliefors version automatically; R needs the function lillie.test.

In the SAS output we see $D = 0.186$ with $p < 0.01$, so we reject H_0 and conclude that GA_est is not normally distributed.

```
proc univariate normal;
   var ga_ultra;
```

```
                  Tests for Normality

Test                   --Statistic---    -----p Value------

Shapiro-Wilk           W     0.698496    Pr < W     <0.0001
Kolmogorov-Smirnov     D     0.185894    Pr > D     <0.0100
Cramer-von Mises       W-Sq  3.729445    Pr > W-Sq  <0.0050
Anderson-Darling       A-Sq  22.17517    Pr > A-Sq  <0.0050
```

Figure 1: Histogram and normal QQ plot for GA by ultrasound



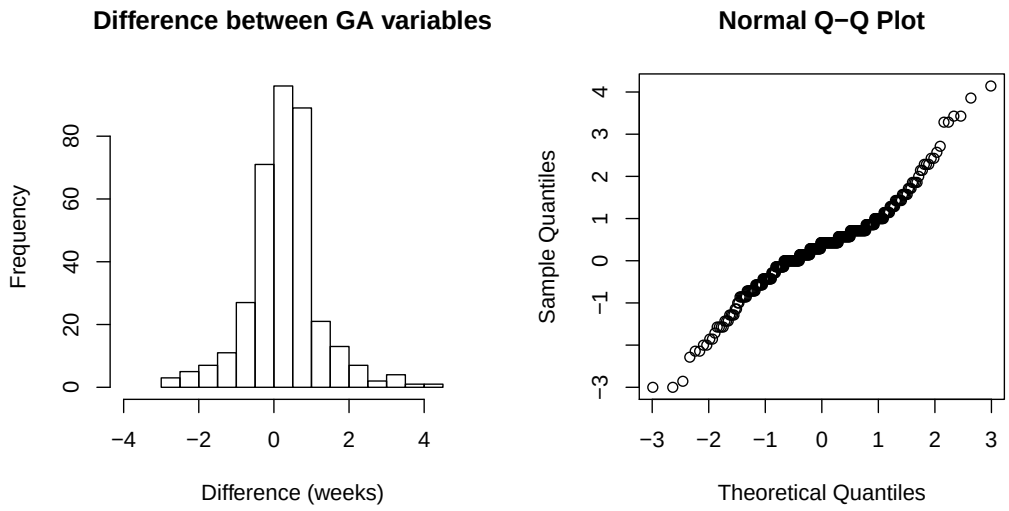
(b) Because the data are paired (each woman has gestational age estimated by each method), we don't have two independent samples. So let $Y_i = X_{1i} - X_{2i}$ where X_{1i} is the GA by ultrasound and X_{2i} the GA estimated at birth for woman i. Assume $E(Y_i) = \mu$ for all i. The observations on different women are independent. We want to test $H_0 : \mu_{\text{diff}} = 0$ versus $H_A : \mu_{\text{diff}} \neq 0$. The histogram in Figure 2 shows that the distribution of differences is reasonably symmetric but the QQ plot suggests that it may not be reasonable to assume normality. The sample size is large, so we will use the CLT / Slutsky's Theorem to give a test using the Z statistic. (Using a t-test would also be reasonable.) The critical region is $C_{0.05} = \{Z : |Z| > 1.96\}$.

$$Z = \frac{\bar{Y} - 0}{s/\sqrt{n}} = \frac{0.3180}{0.960/\sqrt{358}} = 6.268 > 1.96.$$

So we reject H_0 and conclude that the mean GA by ultrasound differs from the mean GA estimated at birth. The sample means are 38.73 and 38.41 months respectively, so the mean GA by ultrasound appears to be larger than the mean GA estimated at birth. A t-test gives similar results (see SAS output below).

```
The TTEST Procedure

Variable:  ga_diff

   N        Mean     Std Dev
 358      0.3180      0.9601

    DF    t Value    Pr > |t|
   357       6.27      <.0001
```

Figure 2: Histogram and normal plot for differences between GA variables



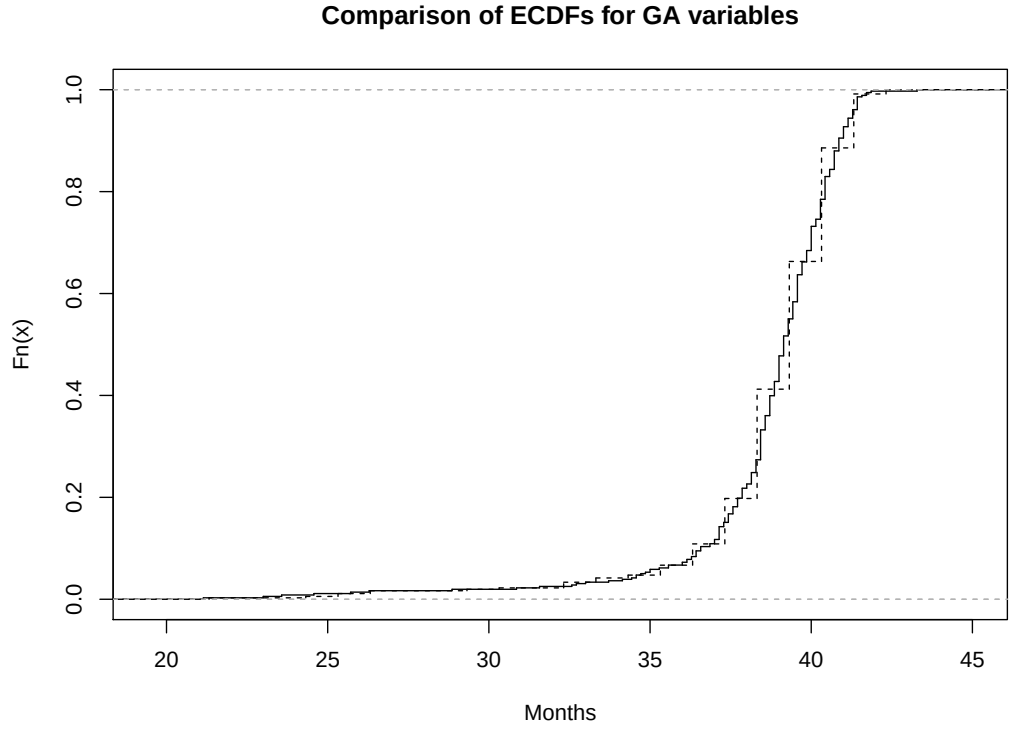
(c) Because the means differ by 0.3180 months, we add this to each GA_est observation. This eliminates the difference between the means. One way to test whether there are other differences between the distributions after eliminating the difference between the means is to use the Kolmogorov-Smirnov test. The KS test in this situation assumes independence of the two samples, but here the two measures are used for each infant and so the assumption of independence is violated. As we don't have a test that does not make the independence assumption we will use the KS test even though it is not ideal.

We want to test $H_0 : F_1(y) = F_2(y)$ for all y versus $H_A : F_1(y) \neq F_2(y)$ for at least one y, where F_1 and F_2 are the CDFs of the two GA variables. To run the KS test we first need to create a single variable that has the GA values of both types, along with an indicator of which are the GA_ultra and which the GA_est values. In the SAS code that follows these are the variables GA and GA_GROUP, respectively. The "mc" in the last line of the code is so that SAS doesn't take forever to run. The p-values for both the asymptotic and the exact versions of the test are < 0.05, so we reject the hypothesis that the two distributions have the same shape. However, looking at the plot in Figure 3, the only noticeable difference is that the step function for the GA_EST version has bigger steps because it is given in whole weeks whereas the other version is in weeks plus days. So, although the difference is statistically significant it isn't really a meaningful difference.

A test for equality of variances for the two variables does not reject the null hypothesis of equality (though here too the assumption of independence of the two samples is not met. The histogram in Figure 4 shows that the distribution of differences is reasonably symmetric.

```
proc npar1way;
   class ga_group;
   var ga;
   exact ks / mc;

  Kolmogorov-Smirnov Two-Sample Test

D = max |F1 - F2|                0.1397
Asymptotic Pr >  D               0.0019

Monte Carlo Estimate
Exact Pr >= D                    0.0010
```

Figure 3: ECDFs for the GA variables after eliminating the difference in means



(d) The following table has the two versions of GA classified as stated. The observed proportion of agreement is $p_a = (33 + 170 + 91)/359 = 0.82$. As seen from the SAS output, the chance-corrected measure of agreement is $\kappa = 0.68$ and the associated 95% CI for the true agreement is $(0.61, 0.75)$. The agreement is reasonable, though not great.

Figure 4: Histogram of the difference between the GA variables after eliminating the difference in means

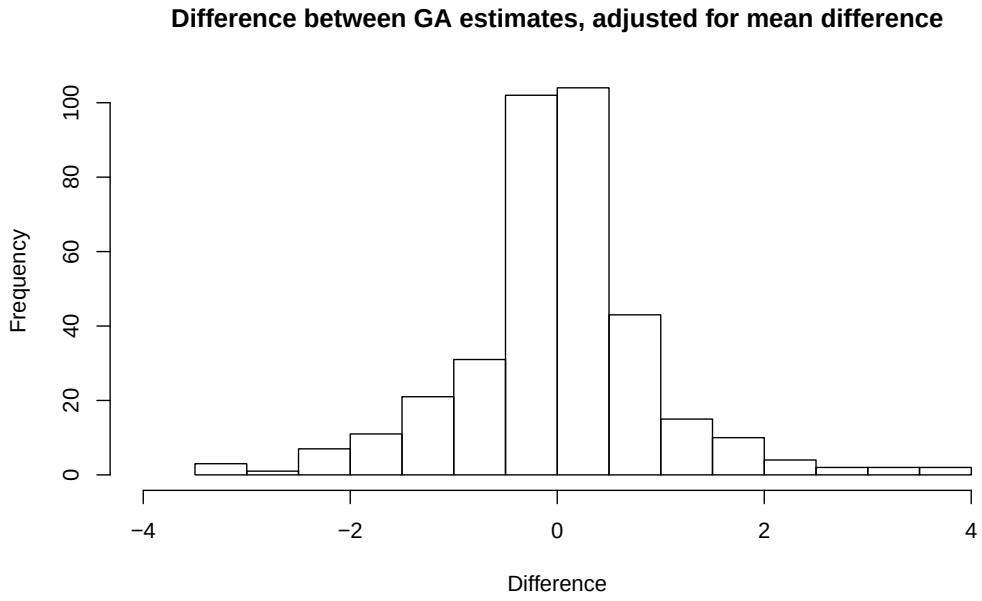


Table of ga_cat by ga_catp

ga_cat	ga_catp			
Frequency	1. LT 37	2. 37-40	3. GE 40	Total
1. LT 37	33	6	0	39
2. 37-40	6	170	30	206
3. GE 40	0	22	91	113
Total	39	198	121	358

Frequency Missing = 1

Kappa Statistics

Statistic	Value	ASE	95% Confidence Limits	
Simple Kappa	0.6826	0.0366	0.6108	0.7544
Weighted Kappa	0.7182	0.0335	0.6526	0.7838

Effective Sample Size = 358
Frequency Missing = 1

(e) If randomization is spread evenly across the calendar year, we would expect the proportion of women randomized in January to be 31/365, the proportion in February to be 28/365, etc. We need to conduct a goodness of fit test to see whether the distribution of the number of women randomized is consistent with this. This is similar to the example starting on page 33 of the overheads on "Categorical Data: Contingency Tables" with 12 probabilities rather than just 3.

$$H_0 : \pi_{\text{Jan}} = 31/365 = 0.084932, \pi_{\text{Feb}} = 28/365 = 0.076712, \ldots \pi_{\text{Dec}} = 31/365 = 0.084932$$

$$H_A : \text{at least one of the equalities is false.}$$

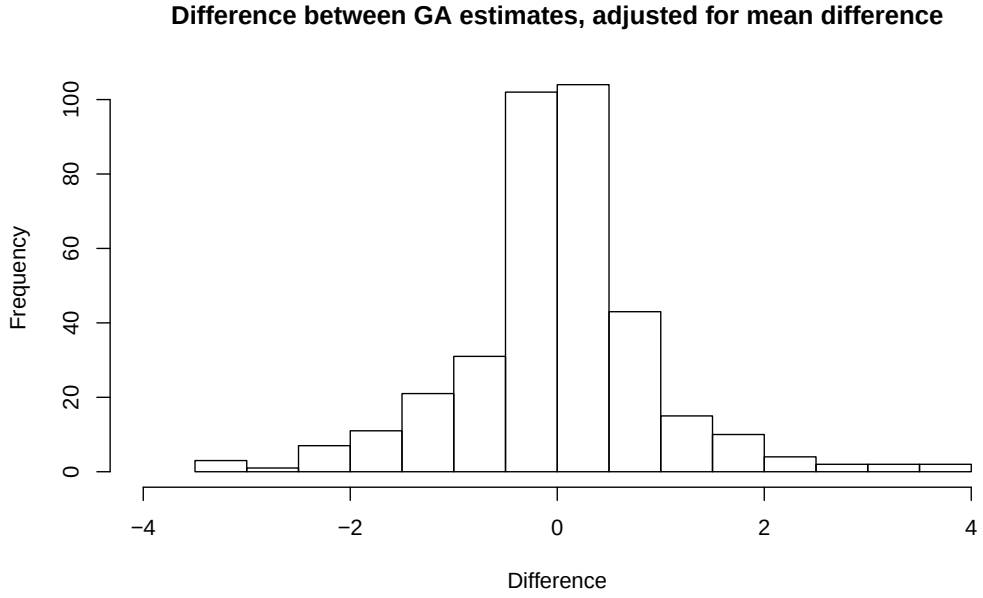
Using the SAS code and output below, $p = 0.0004$, so we reject H_0 and conclude that the proportion of women randomized each month is not consistent with the number of days in the month. Some slight variation from expected is likely to be because of the number of weekend days in a particular month. But we see that December, in particular, has a much lower percent randomized than expected just by the length of the month. Because of holidays in December, which means fewer working days and also potential participants having other priorities than being in a clinical trial, trials often struggle to randomize many participants in that month.

```
proc freq data=bw;
   table rand_month / testp=(8.4932 7.6712 8.4932 8.2192 8.4932 8.2192
        8.4932 8.4932 8.2192 8.4932 8.2192 8.4932);

The FREQ Procedure
                                              Test
rand_month    Frequency     Percent     Percent
-------------------------------------------------
         1           22        6.16        8.49
         2           17        4.76        7.67
         3           42       11.76        8.49
         4           24        6.72        8.22
         5           41       11.48        8.49
         6           35        9.80        8.22
         7           40       11.20        8.49
         8           22        6.16        8.49
         9           32        8.96        8.22
        10           39       10.92        8.49
        11           29        8.12        8.22
        12           14        3.92        8.49

     Chi-Square Test
for Specified Proportions
-------------------------
Chi-Square        33.4334
DF                     11
Pr > ChiSq         0.0004

Effective Sample Size = 357
Frequency Missing = 2
```

(f) Table 1 gives percentages of women randomized by month and live births by month. The percentage of birth each month is much more even than that of the number of women randomized. (I said not to do a test, but if one does the same test as in (e) for the births it yields $p = 0.5$.) That could be because the number of births in a month is not influenced substantially by the number of holidays in the month. (It is probably also because I had an error when I created the dataset – I used the month from each woman's date of birth rather than from her infant's date of birth.)

Month	1	2	3	4	5	6	7	8	9	10	11	12
Randomized	6.2	4.8	11.8	6.7	11.5	9.8	11.2	6.2	9.0	10.9	8.1	3.9
Births	6.1	6.4	8.6	7.5	7.5	9.7	9.5	9.7	8.6	11.4	7.2	7.5

Table 1: Percentages of women randomized each month and births each month



(g) We use a χ^2 test of trend. Let ρ_i denote the probability of preterm delivery in previous pregnancy category i. We want to test $H_0 : \rho_0 = \rho_1 = \rho_2 = \rho_{3+}$ against $H_A : \rho_0 \leq \rho_1 \leq \rho_2 \leq \rho_{3+}$ or $\rho_0 \geq \rho_1 \geq \rho_2 \geq \rho_{3+}$ with at least one of the inequalities being strict. From the SAS output, $X^2_{\text{trend}} = (-0.0335)^2 = 0.0011$ with $p = 0.9733$. So we do not reject H_0 and conclude there isn't much evidence for a monotonic trend in the risk of preterm delivery with the number of previous pregnancies.

```
Table of preterm by ppnum

preterm     ppnum

Frequency|
Col Pct  | 0      | 1      | 2      |3+      |  Total
---------+--------+--------+--------+--------+
       0 |     57 |    117 |     72 |     28 |    274
         |  85.07 |  92.13 |  88.89 |  84.85 |
---------+--------+--------+--------+--------+
       1 |     10 |     10 |      9 |      5 |     34
         |  14.93 |   7.87 |  11.11 |  15.15 |
---------+--------+--------+--------+--------+
Total          67      127       81       33      308

Frequency Missing = 51

Cochran-Armitage Trend Test
---------------------------
Statistic (Z)       -0.0335
One-sided Pr <  Z    0.4867
Two-sided Pr > |Z|   0.9733
```

(h) Now we want to test whether there is an association between treatment group and preterm delivery. We use a χ^2 test of association or, equivalently, test whether the proportion of preterm deliveries is the same in the two treatment groups. We want to test H_0 : preterm delivery is independent of treatment group, versus H_A : preterm delivery is associated with treatment group. The χ^2 test yields $p = 0.14$, hence we do not reject the null hypothesis. From the SAS output we see that the preterm percentages in the two groups are 13.5 and 8.6, that is, the percentage is actually nominally higher in the prenatal treatment group. So there is no evidence that treatment of periodontal disease reduces the risk of preterm delivery.

```
group     preterm

Frequency|
Row Pct  |       0|       1|  Total
---------+--------+--------+
       1 |    148 |     23 |    171
         |  86.55 |  13.45 |
---------+--------+--------+
       2 |    171 |     16 |    187
         |  91.44 |   8.56 |
---------+--------+--------+
Total         319       39      358

Statistics for Table of group by preterm

Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      2.2040    0.1376
```

(i) We have pocket depth measurements on each woman at two time points (apart from some missing values, which we exclude from this analysis). Because the two measurements on each woman are not independent, we cannot use a two-sample test. Instead, we calculate the difference and test whether the mean of the differences is zero, that is, $H_0 : \mu_{\text{diff}} = 0$ versus $H_A : \mu_{\text{diff}} \neq 0$. Taking the difference as the pocket depth after delivery minus the pocket depth at baseline, the difference will be positive if pocket depth has increased (that is, periodontal disease has progressed) and negative if it has declined (that is, if there has been an improvement).

Even after omitting missing values, the sample size is large ($n = 286$), so we can rely on the CLT and Slutsky's Theorem to give a test using the Z statistic.

The critical region is $C_{0.05} = \{Z : |Z| > 1.96\}$.

Using the SAS output below, the test statistic is $z = (-0.0835 - 0)/\sqrt{0.2173/286} = -3.03 > 1.96$. The associated p-value is 0.002. (Using a one-sample t-test gives a very similar p-value – see the SAS output.)

So we reject H_0 and conclude that the mean average pocket depth changed from baseline to delivery. The estimate of mean change is -0.08mm. That is average pocket depth became slightly worse, even though about half of the participants had their periodontal disease treated in the interim.

```
N                          286    Sum Weights               286
Mean                -0.0835129    Sum Observations    -23.884703
Std Deviation       0.46619515    Variance            0.21733791

           Tests for Location: Mu0=0

Test           -Statistic-    -----p Value------

Student's t    t  -3.02949    Pr > |t|    0.0027
```

(j) Now we do have two independent sets of measurements – the data from the two treatment groups (with the data within each group being the difference between the pocket depth measurements at the two time points, as in part (e)). The sample sizes in the two groups are still large ($n_1 = 61$ in the prenatal group and $n_2 = 62$ in the post-partum group), so we can again rely on the CLT and Slutsky's Theorem to give a test using the Z statistic. Here $H_0 : \mu_{\text{diff},1} = \mu_{\text{diff},2}$ versus $H_A : \mu_{\text{diff},1} \neq \mu_{\text{diff},2}$.

The critical region is again $C_{0.05} = \{Z : |Z| > 1.96\}$.

$$Z = \frac{(\bar{Y}_1 - \bar{Y}_2) - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(0.0130 - (-0.1662)) - 0}{\sqrt{\frac{0.4910^2}{132} + \frac{0.4284^2}{154}}} = 3.26$$

As $3.26 > 1.96$ we reject the null hypothesis and conclude there is a difference between the change in pocket depth in the prenatal treatment group compared with the post-partum treatment group. The associated p-value is 0.0011. Using a t-test yields similar results

```
The TTEST Procedure

Variable:  pd_chng

group            N        Mean     Std Dev

1              132      0.0130      0.4910
2              154     -0.1662      0.4284
Diff (1-2)              0.1792      0.4583


group         Method               Mean       95% CL Mean

1                                0.0130     -0.0716   0.0975
2                               -0.1662     -0.2344  -0.0980
Diff (1-2)    Pooled             0.1792      0.0721   0.2862
Diff (1-2)    Satterthwaite      0.1792      0.0710   0.2873

Method           Variances        DF    t Value    Pr > |t|

Pooled           Equal           284       3.30      0.0011
Satterthwaite    Unequal      262.16       3.26      0.0013
```

(k) In part (i) we saw that overall there is a very marginal improvement in the mean pocket depth. From part (j) we see that the change in pocket depth differs significantly between the two treatment groups. Looking at the point estimates and confidence intervals for the change within each group (in the SAS output in part (j), we see that pocket depth tended to get worse in the post-partum treatment group. For the prenatal treatment group the point estimate shows a small improvement in mean pocket depth, but the associated 95% confidence interval includes 0, so there is not a statistically significant (or clinically meaningful improvement. So, even if effective periodontal therapy does have the potential to reduce the risk of premature birth or low birthweight, the periodontal therapy given in this study does not appear to have had much effect on periodontal disease, but is better than leaving the periodontal disease untreated and so may have an effect on birth outcomes.

Question 3

(a) Below is a 2×2 table of exposure by case status along with the column percentages, that is the percentage in each exposure category among cases and among controls.

	Control	Case	Total
Exposed = 0	69	51	120
	79.3	58.6	
Exposed = 1	18	36	54
	20.7	41.4	
Total	87	87	174

Let π_1 be the probability of being exposed in the control group and π_2 the corresponding probability in the case group. We want to test $H_0 : \pi_1 = \pi_2$ versus $H_A : \pi_1 \neq \pi_2$. The sample size is large enough to use a χ^2 test of association. The critical region is $C_{0.05} = \{X^2 : X^2 > \chi^2_{1,0.95} = 3.84\}$. From the SAS output below we see that $X^2 = 8.7 > 3.84$ and the corresponding p-value is 0.003. Thus we reject the null hypothesis and conclude that moderate to severe periodontal disease is associated with premature birth.

```
Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     1      8.7000    0.0032
```

(b) Because this is a case-control study, the appropriate measure of association is the odds ratio. From the 2×2 table we obtain

$$\widehat{\text{OR}} = \frac{n_{11}n_{22}}{n_{21}n_{12}} = \frac{69 \times 36}{51 \times 18} = 2.71.$$

This is confirmed by the SA output below, which gives the 95% CI as $(1.38, 5.30)$.

```
            Estimates of the Common Relative Risk (Row1/Row2)

Type of Study     Method                  Value     95% Confidence Limits
-----------------------------------------------------------------------
Case-Control      Mantel-Haenszel        2.7059       1.3823       5.2967
  (Odds Ratio)    Logit                  2.7059       1.3823       5.2967
```

(c) Using the Mantel-Haenszel method, from the SAS code and output below we obtain an estimated odds ratio (adjusted for age) of 3.05, with 95% CI $(1.48, 6.28)$.

```
proc freq;
   table age_group*exposed*case / norow nopercent cmh;
```

```
          Estimates of the Common Relative Risk (Row1/Row2)

Type of Study     Method                  Value     95% Confidence Limits
-------------------------------------------------------------------------
Case-Control      Mantel-Haenszel        3.0506       1.4819       6.2802
  (Odds Ratio)
```

(d) The adjusted odds ratio of 3.05 is reasonably similar to the unadjusted one of 2.71, so age group does not appear to be a substantial confounder of the association between periodontal disease and premature birth. Investigating whether age group is associated with both the exposure and the outcome will show that numbers of cases and controls are equal within each age group (because of the matching on age), so age and case status are not associated in this dataset. From separate 2×2 tables for the three age groups, we obtain estimated odds ratios of 2.23, 4.86 and 9.63, respectively. This suggests that the odds ratios are not homogeneous across the age groups and so it may not be appropriate to pool them using the Mantel-Haenszel estimator. (We have not covered a test of homogeneity of the odds ratios and I did not expect you to look for such a test. The Breslow-Day test for homogeneity does not reject the null hypothesis of homogeneity. This test is part of the output from the SAS code above.)

```
     Breslow-Day Test for
Homogeneity of the Odds Ratios
------------------------------
Chi-Square              1.8405
DF                           2
Pr > ChiSq              0.3984
```

(e) For this part we need the 2×2 table in a different form. To obtain this in SAS, we need each pair to be a single observation in the SAS dataset. One way to do this is to rename the exposure variables so that those for cases and controls are distinct, split the dataset into two, one consisting of cases, the other of controls, and then merge the two on the part of the ID that is common to the members of a pair. This yields the following table.

		Controls		
		$E = 0$	$E = 1$	Total
Cases	$E = 0$	44	7	51
	$E = 1$	25	11	36
		69	18	87

We want to test $H_0 : \pi_1 = \pi_2$ versus $H_A : \pi_1 \neq \pi_2$. We do so using McNemar's test statistic. Here $n_{12} + n_{21} = 7 + 25 = 32 > 30$, so we can use the χ^2 approximation. The critical region is $C_\alpha = \{\mathrm{M} : \mathrm{M} > \chi^2_{1,0.95}\} = \{\mathrm{M} : \mathrm{M} > 3.84\}$.

$$\mathrm{M} = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} = \frac{(7-25)^2}{7+25} = 10.1 > 3.84.$$

So we reject the null hypothesis and conclude that moderate to severe periodontal disease is associated with premature birth. From the SAS output we see that the associated p-value is 0.0015.

```
Statistic (S)         10.1250
DF                          1
Asymptotic Pr >  S     0.0015
Exact      Pr >= S     0.0021
```

To estimate the odds ratio, whether one uses n_{12}/n_{21} or n_{21}/n_{12} depends on how the table is set up. Here $\widehat{\mathrm{OR}}_\mathrm{M} = n_{21}/n_{12} = 25/7 = 3.57$.

$$\widehat{\mathrm{Var}}(\ln(\widehat{\mathrm{OR}}_\mathrm{M})) \approx \frac{1}{n_{12}} + \frac{1}{n_{21}} = \frac{1}{25} + \frac{1}{7} = 0.183$$

So an approximate 95% CI on the log scale is $\log(3.57) \pm 1.96 \cdot \sqrt{0.183}$, that is $(0.435, 2.111)$. On the original scale this becomes $(e^{0.435}, e^{2.111}) = (1.54, 8.23)$.

(f) The estimate in (d) is most appropriate because it takes into account the matched case-control design. Although part (c) takes into account age group, it assumes frequency matching rather than individual matching and doesn't take into account the second matching factor (number of previous pregnancies).

(g) The estimated odds ratio of 3.57 in part (d) is fairly large, the lower end of the 95% confidence interval is well above 1 and the p-value from McNemar's test is substantially below 0.05, so the evidence from the matched case control study is strong. Evidence from a case-control study is relatively weak, for several reasons including concerns about the representativeness of the controls, potential differences in recall of exposure by cases versus controls, assessment of timing of exposure relative to becoming a case and, as in any observational study, the possibility that there may be unmeasured confounders. In this particular study it should be relatively easy to recruit appropriate controls (women giving birth in the same facility as the cases), exposure recall is not an issue for the main exposure (measured periodontal disease) but the timing of the exposure measurement is (after giving birth, so it is not known when the periodontal disease developed). Also, there may have been unmeasured confounders, such as smoking.

BIOS663 Homework 1
Due Wednesday, Feb 6 in class

To report a test, provide H_0, the test statistic, the degrees of freedom, the p-value, the decision (accept vs. reject H_0), and an interpretation of the decision in terms of the subject matter.

1. (a) Prove or dis-prove (with details) that
$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 8 & -6 \\ 4 & 1 & 7 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 8 & 0 \\ 4 & 1 & -2 \end{bmatrix}$$
have linearly independent columns, respectively.

   (b) Find the eigenvalues and eigenvectors of
$$\mathbf{C} = \begin{bmatrix} 2 & 1 \\ 2 & 4 \end{bmatrix}$$

2. Suppose
$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \sim N(0, \mathbf{\Sigma}) \text{ where } \mathbf{\Sigma} = \begin{bmatrix} 2 & 0 & 0.6 \\ 0 & 2 & 0.5 \\ 0.6 & 0.5 & 1 \end{bmatrix}$$

   (a) Derive the distribution of $3x_1 + x_2 + x_3$.

   (b) Derive the distribution of $(x_1, x_2 \mid x_3 = 3)$.

   (c) Calculate $Cov(x_1 + 2x_2, 3x_2 + x_3)$.

3. Suppose $X_1, \ldots, X_k$ are multivariate normally distributed with $X_i \sim N_n(\mu_i, \Sigma_i)$, $i = 1, \ldots, k$. Further, let $Cov(X_i, X_j) = \Sigma_{ij} (i \neq j)$. Suppose $a_1, \ldots, a_k$ are scalars and define $Y = a_1 X_1 + \ldots + a_k X_k$. Find the distribution of Y.

4. *Weighted least squares* is a modification of standard regression analysis that may be used for a set of data when the assumption of variance homogeneity does not hold. (Assume the responses are independent.) If the ith response is an average of m_i equally variable observations, then $\text{Var}(y_i) = \frac{\sigma^2}{m_i}$. In this case, we have the model $\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\, \boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$, where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{V}$, and
$$\mathbf{V} = \begin{bmatrix} \frac{1}{m_1} & 0 & \ldots & 0 \\ 0 & \frac{1}{m_2} & & 0 \\ \vdots & & \ddots & \\ 0 & \ldots & & \frac{1}{m_n} \end{bmatrix}.$$
The fixed and known positive definite matrix $\mathbf{V}_{n\times n}$ has rank n. The weighted least squares estimator of $\boldsymbol{\beta}$ is given by
$$\widehat{\boldsymbol{\beta}}_W = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

(a) Derive the expectation of $\widehat{\boldsymbol{\beta}}_W$, $E[\widehat{\boldsymbol{\beta}}_W]$.

(b) Derive the covariance matrix of $\widehat{\boldsymbol{\beta}}_W$, $\text{Cov}(\widehat{\boldsymbol{\beta}}_W)$.

(c) Find the exact distribution of $\widehat{\boldsymbol{\beta}}_W$. If it is necessary to make any reasonable further assumptions in order to find the distribution of $\widehat{\boldsymbol{\beta}}_W$, provide them.

(d) Explain why this particular choice of $\mathbf{V}$ makes sense when our responses are averages.

Solution to HW 1

1. a. $A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 8 & -6 \\ 4 & 1 & 7 \end{bmatrix}$

solve for $c_1 \begin{pmatrix} 1 \\ 1 \\ 1 \\ 4 \end{pmatrix} + c_2 \begin{pmatrix} 1 \\ 0 \\ 8 \\ 1 \end{pmatrix} + c_3 \begin{pmatrix} 1 \\ 2 \\ -6 \\ 7 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$

$$\Rightarrow \begin{cases} c_1 + c_2 + c_3 = 0 \\ c_1 = -2c_3 \end{cases} \Rightarrow c_2 = c_3 = -\frac{1}{2} c_1$$

and this could solve

$$\begin{cases} c_1 + 8c_2 - 6c_3 = 0 \\ 4c_1 + c_2 + 7c_3 = 0 \end{cases}$$

$\therefore$ Not linearly independent

actually, $A \cdot \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$

$B = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 2 \\ 1 & 8 & 0 \\ 4 & 1 & -2 \end{bmatrix}$

solve for $c_1 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 4 \end{bmatrix} + c_2 \begin{bmatrix} 1 \\ 0 \\ 8 \\ 1 \end{bmatrix} + c_3 \begin{pmatrix} 1 \\ 2 \\ 0 \\ -2 \end{pmatrix} = 0$

first two equations

$$\Rightarrow c_2 = c_3 = -\frac{1}{2} c_1$$

plug into the 3rd & 4th equations

$$\Rightarrow 3c_1 = 0 \Rightarrow c_1 = c_2 = c_3 = 0$$

$\Rightarrow$ linearly independent

1.b Solve $\left| \begin{bmatrix} 2-\lambda & 1 \\ 2 & 4-\lambda \end{bmatrix} \right| = 0$

$\Rightarrow \lambda = 3 \pm \sqrt{3}$

for $\lambda_1 = 3+\sqrt{3}$ solve v_1 s.t. $\begin{bmatrix} 2-(3+\sqrt{3}) & 1 \\ 2 & 4-(3+\sqrt{3}) \end{bmatrix} v_1 = 0$

$\Rightarrow v_1 = c \begin{pmatrix} \sqrt{3}-1 \\ 2 \end{pmatrix}$

similarly, for $\lambda_2 = 3-\sqrt{3} \Rightarrow v_2 = c \begin{pmatrix} -(\sqrt{3}+1) \\ 2 \end{pmatrix}$

~~1.c. Cov $(X_1 + 2X_2, 3X_2 + X_3)$~~

$=$

2. a. Linear combination of ~~joint~~ multi- normal r.v.'s is still normal; what remains to derive is the mean & variance of that normal distribution

$E[3X_1 + X_2 + X_3]$

$= 3E[X_1] + E[X_2] + E[X_3] = 0$

$Var[3X_1 + X_2 + X_3]$

$= 9Var(X_1) + Var(X_2) + Var(X_3)$

$+ 2 \cdot 3 \cdot Cov(X_1, X_2) + 2Cov(X_2, X_3) + 2 \cdot 3 \cdot Cov(X_1, X_3)$

$= 25.6$

$\Rightarrow 3X_1 + X_2 + X_3 \sim N(0, 25.6)$

2.b. Define $Y_1 = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, $Y_2 = (X_3)$

Then $\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$

where $\mu_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\mu_2 = (0)$

$\Sigma_{11} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$, $\Sigma_{12} = \begin{pmatrix} 0.6 \\ 0.5 \end{pmatrix}$, $\Sigma_{21} = \Sigma_{12}^T$, $\Sigma_{22} = (1)$

Using the formula

$$Y_1 \mid Y_2 = b \sim N\left(\mu_1 + \Sigma_{12} \cdot \Sigma_{22}^{-1} (b - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \right)$$

with $b = 3$

$$\Rightarrow Y_1 \mid Y_2 = 3 \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 0.6 \\ 0.5 \end{pmatrix} (1)^{-1} ((3) - (0)) , \right.$$

$$\left. \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} - \begin{bmatrix} 0.6 \\ 0.5 \end{bmatrix} [1]^{-1} [0.6 \ 0.5] \right)$$

$$\Rightarrow Y_1 \mid Y_2 = 3 \sim N\left(\begin{pmatrix} 1.8 \\ 1.5 \end{pmatrix}, \begin{pmatrix} 1.64 & -0.30 \\ -0.30 & 1.75 \end{pmatrix} \right)$$

2.c. $\mathrm{Cov}(X_1 + 2X_2, \ 3X_2 + X_3)$

$= \mathrm{Cov}(X_1, 3X_2) + \mathrm{Cov}(2X_2, 3X_2) + \mathrm{Cov}(2X_2, X_3) + \mathrm{Cov}(X_1, X_3)$

$= 3 \cdot 0 + 6 \cdot 2 + 2 \cdot 0.5 + 0.6$

$= 13.6$

3. Any linear combination of multi-normal r.v.'s is still normal with the mean

$$E[Y] = E\left[\sum_{i=1}^{k} a_i X_i\right]$$

$$= \sum_{i=1}^{k} a_i E[X_i]$$

$$= \sum_{i=1}^{k} a_i \mu_i$$

and variance

$$Var[Y] = Var\left[\sum_{i=1}^{k} a_i X_i\right]$$

$$= \sum_{i=1}^{k} Var[a_i X_i] + \sum_{i=1}^{k} \sum_{j \neq i, j=1}^{k} Cov[a_i X_i, a_j X_j]$$

$$= \sum_{i=1}^{k} a_i^2 Var(X_i) + 2\sum_{i=1}^{k} \sum_{j=1, j \neq i}^{k} a_i a_j Cov(X_i, X_j)$$

$$= \sum_{i=1}^{k} a_i^2 \Sigma_{ii} + \sum_{i=1}^{k} \sum_{j=1, j \neq i}^{k} a_i a_j \Sigma_{ij}$$

$$\therefore \sum_{1}^{k} a_i X_i \sim N\left(\sum_{1}^{k} a_i \mu_i + \sum_{1}^{k} a_i^2 \Sigma_{ii} + \sum_{i=1}^{k} \sum_{j=1, j \neq i}^{k} a_i a_j \Sigma_{ij}\right)$$

4. (a)

$$E[\hat{\beta}_W]$$

$$= E[(X^T V^{-1} X)^{-1} X^T V^{-1} Y]$$

$$= (X^T V^{-1} X)^{-1} X^T V^{-1} E[Y]$$

$$= (X^T V^{-1} X)^{-1} X^T V^{-1} X\beta$$

$$= \beta$$

because $E[Y] = E[X\beta + \varepsilon]$

$= X\beta + E(\varepsilon)$

$= X\beta$

(b) $Cov(\hat{\beta}_W)$

$$= Cov((X^T V^{-1} X)^{-1} X^T V^{-1} Y)$$

$$= (X^T V^{-1} X)^{-1} X^T V^{-1} \, Cov(Y) \, V^{-1} X (X^T V^{-1} X)^{-1}$$

$(\because Cov(Y) = Var(X\beta + \varepsilon) = Var(\varepsilon) = \sigma^2 V)$

$$= (X^T V^{-1} X)^{-1} X^T V^{-1} (\sigma^2 V) V^{-1} X (X^T V^{-1} X)^{-1}$$

$$= \sigma^2 (X^T V^{-1} X)^{-1}$$

(c) If ε is multi-normal, then we immediately know that $\hat{\beta}_W \sim N(\beta, \sigma^2 (X^T V^{-1} X)^{-1})$

Otherwise, only the mean & variance don't give enough information on the exact distribution of $\hat{\beta}_W$

(d) The variance of the average of m_i equally variable observations could be calculated as

$$\begin{aligned} \text{Var}(y_i) &= \text{Var}\left(\sum_{i=1}^{m_i} X_i / m_i\right) \\ &= \left(\frac{1}{m_i}\right)^2 \sum_{i=1}^{m_i} \sigma^2 \\ &= \frac{\sigma^2}{m_i} \end{aligned}$$

Therefore the covariance matrix $\text{Cov}(Y) = \sigma^2 V$

where $V = \begin{bmatrix} \frac{1}{m_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{m_n} \end{bmatrix}$

In other words

- all off-diagnal entries = 0 because we assume the observations are independent
- Diagonal terms are inverse-weighted by the number of observations as more observations mean more information on that y_i => less variable

BIOS663 Homework 2
Due Wednesday, Feb 20 in class.

1. Consider a simple linear regression $Y = X\beta + \epsilon$ with an intercept and one predictor based on a sample of size 4. Or specifically,

$$\begin{pmatrix} 0.5 \\ -0.5 \\ 0.3 \\ 1.2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0.5 \\ 1 & 2 \end{pmatrix} + \epsilon. \tag{1}$$

   Calculate $(X'X)^{-1}$, $X'Y$, $\hat{\beta}$, $\hat{y}$ and $\hat{\epsilon}$ by hand.

2. Consider the model $\mathbf{y} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \beta_4\mathbf{x}_4 + \boldsymbol{\varepsilon}$. Give the appropriate $\mathbf{C}$ and $\boldsymbol{\theta}_0$ for testing the following hypotheses.

   (a) $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$

   (b) $H_0 : \begin{pmatrix} \beta_1 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \beta_2 + 2 \\ \beta_4 \end{pmatrix}$

   (c) $H_0 : \begin{pmatrix} \beta_1 - 2\beta_2 \\ \beta_1 + 2\beta_2 \end{pmatrix} = \begin{pmatrix} 4\beta_3 \\ -6 \end{pmatrix}$

3. Consider the model $\mathbf{y}_{5\times1} = \mathbf{X}_{5\times3}\boldsymbol{\beta}_{3\times1} + \boldsymbol{\varepsilon}_{5\times1}$, where

$$\mathbf{y} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \\ 1 \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{J} & \mathbf{x}_1 & \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}, \text{ and } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix},$$

   with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

   (a) Show as rigorously as possible whether $\theta_1 = \beta_2$ is estimable.

   (b) Show as rigorously as possible whether $\boldsymbol{\theta}_2 = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix}$ is testable.

4. A group of subjects was recruited to a weight loss study in a medical center. The data consist of their weights (y = WGHT), average daily exercise times (x = TIME). One of the objectives in this study is to investigate the effect of TIME on weight loss.

   (a) A partial ANOVA table for estimating WGHT from TIME is given below. Complete the table.

```
Dependent Variable: WGHT

                            Sum of
Source              DF      Squares    Mean Square   F Value    Pr > F

Model               1     2624.670184

Error               96    1827.099916

Corrected Total     97    4451.7701
```

(b) State the model assumptions based on which the ANOVA table was computed.

(c) Is average daily exercise time a significant predictor for predicting weight loss? State your answers in terms of the model from the previous questions and the statistical test of hypothesis.

5. An investigator studied the ozone levels in the South Coast Air Basin of California for the years 1976-1991. He believes that the number of days the ozone levels exceeded 0.20 ppm (the response) depends on the seasonal meteorological index, which is the seasonal average temperature in degrees Celsius (the predictor). The data *hw2.dat*, is provided on Sakai.

   (a) Fit a regression model with the number of high ozone days as the response and the meteorological index as a covariate, and provide estimates of β_0, β_1, their standard errors, and their interpretations.

   (b) Are all of the β's estimable? Why or why not?

   (c) Report a test of the hypothesis that the number of high ozone days is associated with the meteorological index.

   (d) Using the framework of the linear model, report an $\alpha = 0.05$ test of the hypothesis that a 1 degree increase in average temperature is associated with a 12 day increase in the number of days the ozone levels exceed 0.20 ppm.

   (e) Calculate the 95% confidence interval and prediction interval for the expected number of days the ozone level exceeded 0.2 ppm when the seasonal meteorological index is 16.

1) Consider a simple linear regression $Y = X\beta + \epsilon$ with an intercept and one predictor based on a sample of size 4. Or specifically,

$$\begin{bmatrix} 0.5 \\ -0.5 \\ 0.3 \\ 1.2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0.5 \\ 1 & 2 \end{bmatrix} + \epsilon. \tag{1}$$

Calculate $(X'X)^{-1}$, $X'Y$, $\hat{\beta}$, $\hat{y}$, and $\hat{\epsilon}$ by hand.

$$Y = \begin{bmatrix} 0.5 \\ -0.5 \\ 0.3 \\ 1.2 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0.5 \\ 1 & 2 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0.5 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0.5 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 1+1+1+1 & 1+1+0.5+2 \\ 1+1+0.5+2 & 1+1+0.25+4 \end{bmatrix} = \begin{bmatrix} 4 & 4.5 \\ 4.5 & 6.25 \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{|4(6.25) - 4.5(4.5)|} \begin{bmatrix} 6.25 & -4.5 \\ -4.5 & 4 \end{bmatrix} = \frac{1}{4.75} \begin{bmatrix} 6.25 & -4.5 \\ -4.5 & 4 \end{bmatrix} = \begin{bmatrix} 25/19 & -18/19 \\ -18/19 & 16/19 \end{bmatrix}$$
$$\approx \begin{bmatrix} 1.3158 & -0.9474 \\ -0.9474 & 0.8421 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0.5 & 2 \end{bmatrix} \begin{bmatrix} 0.5 \\ -0.5 \\ 0.3 \\ 1.2 \end{bmatrix} = \begin{bmatrix} 0.5 - 0.5 + 0.3 + 1.2 \\ 0.5 - 0.5 + 0.5 * 0.3 + 2 * 1.2 \end{bmatrix} = \begin{bmatrix} 1.5 \\ 2.55 \end{bmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'Y = \begin{bmatrix} 25/19 & -18/19 \\ -18/19 & 16/19 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2.55 \end{bmatrix} = \begin{bmatrix} -42/95 \\ 69/95 \end{bmatrix} \approx \begin{bmatrix} -0.4421 \\ 0.7263 \end{bmatrix}$$

$$\hat{y} = X\hat{\beta} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0.5 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} -42/95 \\ 69/95 \end{bmatrix} = \begin{bmatrix} 27/95 \\ 27/95 \\ -3/38 \\ 96/95 \end{bmatrix} \approx \begin{bmatrix} 0.2842 \\ 0.2842 \\ -0.0789 \\ 1.0105 \end{bmatrix}$$

$$\hat{\epsilon} = (y - \hat{y}) = \begin{bmatrix} 0.5 \\ -0.5 \\ 0.3 \\ 1.2 \end{bmatrix} - \begin{bmatrix} 27/95 \\ 27/95 \\ -3/38 \\ 96/95 \end{bmatrix} = \begin{bmatrix} 41/190 \\ -149/190 \\ 36/95 \\ 18/95 \end{bmatrix} \approx \begin{bmatrix} 0.2158 \\ -0.7842 \\ 0.3789 \\ 0.1895 \end{bmatrix}$$

2) Consider the model $\boldsymbol{y} = \beta_0 + \beta_1\boldsymbol{x_1} + \beta_2\boldsymbol{x_2} + \beta_3\boldsymbol{x_3} + \beta_4\boldsymbol{x_4} + \boldsymbol{\epsilon}$. Give the appreciate $\boldsymbol{C}$ and $\boldsymbol{\theta}_0$ for testing the following hypotheses.

(a) $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 \equiv \begin{matrix}\beta_1 - \beta_4 = 0 \\ \beta_2 - \beta_4 = 0 \\ \beta_3 - \beta_4 = 0\end{matrix}$

$$\boldsymbol{C\beta} = \boldsymbol{\theta_0} \quad \Rightarrow \quad \begin{bmatrix}0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1\end{bmatrix}_{3x5}\begin{bmatrix}\beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4\end{bmatrix}_{5x1} = \begin{bmatrix}\beta_1 - \beta_4 \\ \beta_2 - \beta_4 \\ \beta_3 - \beta_4\end{bmatrix}_{3x1} = \begin{bmatrix}0 \\ 0 \\ 0\end{bmatrix}_{3x1}$$

$$\boldsymbol{C} = \begin{bmatrix}0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 1 & -1\end{bmatrix}_{3x5} \quad \boldsymbol{\theta_0} = \begin{bmatrix}0 \\ 0 \\ 0\end{bmatrix}_{3x1}$$

(b) $H_0 : \begin{pmatrix}\beta_1 \\ \beta_3\end{pmatrix} = \begin{pmatrix}\beta_2 + 2 \\ \beta_4\end{pmatrix}$

$$\boldsymbol{C\beta} = \boldsymbol{\theta_0} \quad \Rightarrow \quad \begin{bmatrix}0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1\end{bmatrix}_{2x5}\begin{bmatrix}\beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4\end{bmatrix}_{5x1} = \begin{bmatrix}\beta_1 - \beta_2 \\ \beta_3 - \beta_4\end{bmatrix}_{2x1} = \begin{bmatrix}2 \\ 0\end{bmatrix}_{2x1}$$

$$\boldsymbol{C} = \begin{bmatrix}0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1\end{bmatrix}_{2x5} \quad \boldsymbol{\theta_0} = \begin{bmatrix}2 \\ 0\end{bmatrix}_{2x1}$$

(c) $H_0 : \begin{pmatrix}\beta_1 - 2\beta_2 \\ \beta_1 + 2\beta_2\end{pmatrix} = \begin{pmatrix}4\beta_3 \\ -6\end{pmatrix}$

$$\boldsymbol{C\beta} = \boldsymbol{\theta_0} \quad \Rightarrow \begin{bmatrix}0 & 1 & -2 & -4 & 0 \\ 0 & 1 & 2 & 0 & 0\end{bmatrix}_{2x5}\begin{bmatrix}\beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4\end{bmatrix}_{5x1} = \begin{bmatrix}\beta_1 - 2\beta_2 - 4\beta_3 \\ \beta_1 + 2\beta_2\end{bmatrix}_{2x1} = \begin{bmatrix}0 \\ -6\end{bmatrix}_{2x1}$$

$$\boldsymbol{C} = \begin{bmatrix}0 & 1 & -2 & -4 & 0 \\ 0 & 1 & 2 & 0 & 0\end{bmatrix}_{2x5} \quad \boldsymbol{\theta_0} = \begin{bmatrix}0 \\ -6\end{bmatrix}_{2x1}$$

3) Consider the model $\boldsymbol{y}_{5x1} = \boldsymbol{X}_{5x3}\boldsymbol{\beta}_{3x1} + \boldsymbol{\varepsilon}_{5x1}$, where

$$\boldsymbol{y} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \\ 1 \end{bmatrix}, \boldsymbol{X} = [\boldsymbol{J} \quad \boldsymbol{x_1} \quad \boldsymbol{x_2}] = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}, \text{ and } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \text{ with } \boldsymbol{\varepsilon} \sim N_5(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

(a) Show as rigorously as possible whether $\theta_1 = \beta_2$ is estimable.

$\boldsymbol{X}$ is not full rank. $r(\boldsymbol{X}) = 2 < 3$ $\qquad\qquad$ $\boldsymbol{x_2} = \boldsymbol{J} - \boldsymbol{x_1}$

$\theta_1 = \beta_2$ is estimable if there exists a $\boldsymbol{T}$ matrix for $\boldsymbol{\theta} = \boldsymbol{C\beta} = \boldsymbol{T}E(\boldsymbol{Y})$such that $\boldsymbol{C} = \boldsymbol{TX}$.

$$\theta_1 = \beta_2 \quad \Rightarrow \quad \boldsymbol{\theta} = \boldsymbol{C\beta} \equiv \theta_1 = [0 \quad 0 \quad 1]\boldsymbol{\beta} = \beta_2$$

$$\Rightarrow \quad \boldsymbol{C} = \boldsymbol{TX} \equiv [0 \quad 0 \quad 1] = [t_1 \quad t_2 \quad t_3 \quad t_4 \quad t_5]\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}$$

$$\Rightarrow \quad [0 \quad 0 \quad 1] = \left[\sum_{i=1}^{5} t_i \quad t_1 + t_2 + t_5 \quad -t_3 - t_4\right]$$

$$\Rightarrow \quad \begin{matrix} t_1 + t_2 + t_3 + t_4 + t_5 = 0 \\ t_1 + t_2 + t_5 = 0 \\ -t_3 - t_4 = 1 \end{matrix} \Rightarrow \begin{matrix} t_3 + t_4 = 0 \\ -t_3 - t_4 = 1 \equiv t_3 + t_4 = -1 \end{matrix}$$

Since $t_3 + t_4 = -1 \neq 0$, there is no $\boldsymbol{T}$ that can satisfy the equation $\boldsymbol{C} = \boldsymbol{TX}$.

$\therefore \theta_1 = \beta_2$ is not estimable

3) Consider the model $\boldsymbol{y}_{5x1} = \boldsymbol{X}_{5x3}\boldsymbol{\beta}_{3x1} + \boldsymbol{\varepsilon}_{5x1}$, where

$$\boldsymbol{y} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \\ 1 \end{bmatrix}, \boldsymbol{X} = [\boldsymbol{J} \quad \boldsymbol{x_1} \quad \boldsymbol{x_2}] = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}, \text{ and } \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \text{ with } \boldsymbol{\varepsilon} \sim N_5(\boldsymbol{0}, \sigma^2\boldsymbol{I}).$$

(b) Show as rigorously as possible whether $\boldsymbol{\theta_2} = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix}$ is testable.

For $\boldsymbol{\theta_2} = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix}$ to be testable, it must also be estimable with either $\boldsymbol{C}$ or $\boldsymbol{M}$ being full rank.

$$\boldsymbol{\theta_2} = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix} \qquad \Rightarrow \qquad \boldsymbol{\theta} = \boldsymbol{C\beta} \equiv \boldsymbol{\theta_2} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix}\boldsymbol{\beta} = \begin{bmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{bmatrix}$$

$$\Rightarrow \qquad \boldsymbol{C} = \boldsymbol{TX} \equiv \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & t_{13} & t_{14} & t_{15} \\ t_{21} & t_{22} & t_{23} & t_{24} & t_{25} \end{bmatrix}\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{bmatrix}$$

$$\Rightarrow \qquad \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{5} t_{1i} & t_{11} + t_{12} + t_{15} & -t_{13} - t_{14} \\ \sum_{i=1}^{5} t_{2i} & t_{21} + t_{22} + t_{25} & -t_{23} - t_{24} \end{bmatrix}$$

These equations are satisfied by $\boldsymbol{T} = \begin{bmatrix} 1/3 & 1/3 & 0 & 0 & 1/3 \\ 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}$. This is one of many solutions for $\boldsymbol{T}$.

$\therefore \boldsymbol{\theta_2} = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix}$ is estimable

$\boldsymbol{C} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \sim \begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & -1 \end{bmatrix} \Rightarrow r(\boldsymbol{C}) = 2$ $\therefore \boldsymbol{C}$ is full rank

Since $\boldsymbol{\theta_2} = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix}$ is estimable and $\boldsymbol{C}$ is full rank, $\boldsymbol{\theta_2} = \begin{pmatrix} \beta_0 + \beta_1 \\ \beta_0 - \beta_2 \end{pmatrix}$ is testable.

4) A group of subjects was recruited to a weight loss study in a medical center. The data consist of their weights (y = WGHT), average daily exercise time (x = TIME). One of the objectives in this study is to investigate the effect of TIME on weight loss.

(a) A partial ANOVA table for estimating WGHT from TIME is given below. Complete this table.

Dependent Variable: WGHT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2624.670184	2624.670184/1 = 2624.670184	137.906162 ~F(1,96)	2.86E-20
Error	96	1827.099916	1827.099916/96 = 19.032291		
Corrected Total	97	4451.7701			

(b) State the model assumptions based on which the ANOVA table was computed.

HILE-Gauss:

1. Homogeneity Assumption: We assume each row of ε has same variance σ^2.
2. Independence Assumption: We assume each row of ε is statistically independent of every other row.
3. Linearity Assumption: We assume the expected value of the response are linear functions of the parameter. $E(y) = X\beta$.
4. Existence Assumption: We observe values of random variables with finite variance. $H_0: \sigma^2_{model} = \sigma^2_{error}$
5. The error term follows a Gaussian distribution. $\varepsilon_i \sim N(0, \sigma^2)$

(c) Is average daily exercise time a significant predictor for predicting weight loss? State your answers in terms of the model from the previous questions and the statistical test of hypothesis.

Yes, daily exercise time is a significant predictor for predicting weight loss.
The test statistic is 137.9062.
The F-test of $\beta_{WGHT} = 0$ for $y = X\beta + \varepsilon$ generates a p-value < 0.0001.
We reject the null hypothesis that $\beta_{WGHT} = 0$ (the average daily exercise time is not significant).
Therefore, daily exercise time appears to be a significant predictor for predicting weight loss.

5) An investigator studied the ozone levels in the South Coast Air Basin of California for the years 1976-1991. He believes that the number of days the ozone levels exceeded 0.20 ppm (the response) depends on the seasonal meteorological index, which is the seasonal average temperature in degrees Celsius (the predictor). The data *hw2.dat*, is provided on Sakai.

(a) Fit a regression model with the number of high ozone days as the response and the meteorological index as a covariate and provide estimates of β_0, β_1, their standard errors, and their interpretations.

Table 5.1: Regression Model for Number of High Ozone Days

Variable	Parameter Estimate	Standard Error	Interpretation
β_0	-192.98	163.503	The number of high ozone days with a meteorological index temperature of 0.
β_1	15.30	9.421	The change in the number of days with high ozone with every increase in the meteorological index temperature by 1 degree Celsius.

(b) Are all of the β's estimable? Why or why not?

Yes. X is full rank ($r(\boldsymbol{X}) = 2 = p = r$). Therefore, all the β's are estimable.

(c) Report a test of the hypothesis that the number of high ozone days is associated with the meteorological index.

Hypothesis: $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$

Test Statistic: $t_{obs} = \frac{\widehat{\beta}_1 - \beta_1}{SE_{\beta_1}} = \frac{15.30-0}{9.421} = 1.62 \sim t_{14}$

Degrees of Freedom: $df = 16 - 2 = 14$

P-value: $\Pr(|t| > 1.62) = 2 * (1 - \Pr(t \leq 1.62)) = 0.1267$

Decision: We fail to reject the null hypothesis.

Interpretation: There is insufficient evidence to suggest that there is an association between the number of high ozone days and meteorological index.

(d) Using the framework of the linear model, report an $\alpha = 0.05$ test of the hypothesis that a 1 degree increase in average temperature is associated with a 12 day increase in the number of days the ozone levels exceed 0.20 ppm.

Hypothesis: $H_0: \beta_1 = 12$ vs. $H_1: \beta_1 \neq 12$

Test Statistic: $t_{obs} = \frac{\hat{\beta}_1 - \beta_1}{SE_{\beta_1}} = \frac{15.30-12}{9.421} = 0.35 \sim t_{14}$ or
$F = 0.12 \sim F(1,14)$

Degrees of Freedom: $df = 16 - 2 = 14$

Critical Region: $C_\alpha = \left\{t: |t| > t_{v,1-\frac{\alpha}{2}}\right\} \rightarrow C_{0.05} = \{t: |t| > 2.1448\}$

P-value: $\Pr(|t| > 0.35) = 2 * (1 - \Pr(t \leq 0.35)) = 0.7316$ or

$\Pr(F_{1,14} > 0.12) = 1 - \Pr(F_{1,14} \leq 0.12) = 0.7316$

Decision: We fail to reject the null hypothesis.

Interpretation: There is insufficient evidence to suggest that a 1 degree increase in average temperature is not associated with a 12 day increase in the number of days the ozone levels exceed 0.20 ppm.

(e) Calculate the 95% confidence interval and prediction interval for the expected number of days the ozone level exceeded 0.2 ppm when the seasonal meteorological index is 16.

95% Confidence Interval: $(21.7579, 81.7581)$

When the seasonal meteorological index is 16, there is a 95% confidence that the average number of days the ozone level exceeded 0.2 ppm is between 21.76 and 81.76 days.

95% Prediction Interval: $(-7.4416, 110.9576)$

Based on the observed data, there is a 95% chance that a seasonal meteorological index of 16 will result in between 0 and 110.96 days that the ozone level exceeded 0.2ppm.

BIOS663 Homework 3
Due noon on Tuesday, March 5 to my mailbox.

1. A group of subjects was recruited to a weight loss study in a medical center. The data consist of their weights (y = WGHT), average daily exercise times (x1 = TIME) and average daily running mileages (x2 = RUN). One of the objectives in this study is to investigate the effect of x1 and x2 on weight loss.

   (a) A partial ANOVA table for estimating WGHT from TIME is given below. Complete the table.

   Dependent Variable: WGHT

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2624.670184			
Error	96	1827.099916			
Corrected Total	97	4451.7701			

   (b) State the model assumptions based on which the ANOVA table was computed.

   (c) Is average daily exercise time a significant predictor for predicting weight loss? State your answers in terms of the model from the previous questions and the statistical test of hypothesis.

   (d) To further explore the unexplained variation in the data, and in order to improve the predictive power of the model, the average daily running mileage (RUN) is also considered. The result is summarized in the following table. Does the analysis suggest that neither variable is significant? Why and why not?

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TIME	1	88.173	88.173	3.10	0.08
RUN	1	70.339	70.339	2.47	0.12

2. A data set was collected by the Environmental Protection Agency (EPA) at the Health Effects Research Laboratory at UNC: Chapel Hill. One hundred seventy-two young adult males received a battery of pulmonary function tests. (The data are described in more detail in Muller and Fetterman on page 536).

For this homework, fit a model with average forced vital capacity (FVC) (in ml) as the outcome and height, weight, body mass index (BMI=$\frac{weight(kg)}{(height(m))^2}$), age, average treadmill elevation, average treadmill speed, temperature, barometric pressure, and humidity as predictors. For the purpose of added-in-order tests, assume that this order is the preferred order for testing. The data are available on the course website in FILEN.DAT with associated SAS file hw3.SAS.

To report a test, provide H_0, the test statistic, the degrees of freedom, the p-value, the decision (accept/reject H_0), and an interpretation of the result in terms of the subject matter.

(a) Use Proc GLM to produce a table like Table 4.8.1 in Muller and Fetterman (pg56) (details about the table can be found in Lecture7.pdf, pg29) with the following predictors: height, weight and age. The table should contain six df values, six SS values, four MS values, three F values, and three p-values.

(b) Report the test of whether the group of predictors (height, weight, body mass index, age, average treadmill elevation, average treadmill speed, temperature, barometric pressure, and humidity) is important.

(c) Report the corrected R^2 for these data.

(d) Give the two models being compared in testing the following hypotheses for these data, and report each test.

  i. H_1: The entire group of predictors provides no useful information about FVC.

  ii. H_2: Height provides no information about FVC, not adjusting for effects of other predictors (i.e., in a simple regression model).

  iii. H_3: Adjusting for weight and BMI, height does not provide any additional information about FVC.

  iv. H_4: After adjusting for weight, BMI, age, elevation, speed, temperature, barometric pressure, and humidity, height does not provide any additional information about FVC.

  v. H_5: The group of body size variables (height, weight, BMI) provides no additional information about FVC compared to a model for only the mean level of FVC.

  vi. H_6: The group of body size variables (height, weight, BMI) provides no additional information about FVC after adjusting for age, elevation, speed, temperature, barometric pressure, and humidity.

(e) Report a test of the hypothesis that humidity has no affect on FVC after adjusting for all the other variables in the model.

(f) Describe the relationship between the body size variables and FVC in these data.

(g) Based on the original model, which characteristics are associated with the best (largest) FVC?

3. For the same data in Q2, consider the following model

$$\begin{aligned} FVC_i &= \beta_0 + \beta_1 HEIGHT_i + \beta_2 WEIGHT_i + \beta_3 BMI_i + \beta_4 AREA_i \\ &\quad + \beta_5 AGE_i + \beta_6 AVTREL_i + \beta_7 AVTRSP_i + \beta_8 AVTREL_i AVTRSP_i \\ &\quad + \beta_9 TEMP_i + \beta_{10} BARM_i + \beta_{11} HUM_i + \varepsilon_i, \end{aligned}$$

$i = 1, \ldots, n$.

   (a) Compute the following correlations, giving the interpretation of each, between FVC and age, and report tests of the hypotheses that each correlation equals zero.

      i. the correlation between age and FVC, controlling both for all the other variables in the model

      ii. the correlation between age and FVC, controlling only age for all the other variables in the model

      iii. the simple correlation between age and FVC (not controlling for any other variables)

   (b) Provide and interpret the following diagnostics (include subject ID when appropriate) for the regression model.

      i. Largest 5 studentized residuals (in absolute value)

      ii. Results of a test of the Gaussian distribution for the studentized residuals

      iii. Histogram of the studentized residuals

      iv. Plot of studentized residuals versus predicted values

1.

(a)

Dependent Variable: WGHT					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2624.670184	2624.670184	137.91	<0.001
Error	96	1827.099916	19.032291		
Corrected total	97	4451.7701			

(b)

The model assumptions behind the model $WGHT = \beta_0 + \beta_1 * TIME + \varepsilon$ are:

(1) existence assumption: assume ε_i has finite first and second moment. In other words, we observe values of random variables with finite variance.

(2) linearity assumption: we assume that the expected values of the weight (WGHT) are linear functions of the average daily exercise times (TIME).

(3) independent assumption: we assume that each element of ε is statistically independent of every other.

(4) homogeneity assumption: we assume that each element of ε has the same variance σ^2.

(5) Gaussian error assumption: we assume that $\varepsilon_i \sim N(0, \sigma^2)$. Note that the normality assumption is needed for the F test, but not needed for the estimates.

(c)

The average daily exercise time is a significant predictor for predicting weight loss. In this specific hypothesis testing, the null hypothesis is that the average daily exercise time is not a significant predictor for predicting weight loss (H_0: $\beta_1 = 0$); and the alternative hypothesis is that the average daily exercise time is a significant predictor for predicting weight loss (H_A: $\beta_1 \neq 0$). According to the ANOVA table, the test statistic $F_{obs} = 137.91$, which follows F distribution with degree of freedom of 1 and 96. The corresponding p-value is less than 0.05. Therefore, we reject the null hypothesis. The result can be interpreted as the following: assuming that the average daily exercise time is not a significant predictor for predicting weight loss, then the probability of

observing the data that are as extreme as ours or more extreme is less than 0.05, which is too small for us to believe that the null hypothesis is true.

(d)

The analysis does not suggest that neither variable is significant, essentially because they are correlated covariates so that the addition of one additional covariate does not provide additional significant information, given the other covariate is in the model. More specifically, the output is for the type III test, which refers to the statistical significance of the added-last test, given that all the other variables are in the model. In other words, the results are interpreted as: given the RUN is in the model, then the p-value for the test of the regression coefficient of TIME after being added to the model is 0.08; and given the TIME is in the model, then the p-value for the test of the regression coefficient of RUN after being added to the model is 0.12.

However, if we add either TIME or RUN to the intercept only model of WGHT, it is still possible that they are significant variables.

7

a.

Source	Df	SS	MS	Fobs	p
Intercept	1.0	4839362527.0	4839362527.0	11694.6	<.0001
Model (Un.)	4.0	4885896692.4	1221474173.1	2951.8	<.0001
Model (Cor.)	3.0	46534165.4	15511388.5	37.5	<.0001
Error (Res.)	166.0	68692765.6	413811.8		
Total (Un.)	170.0	4954589458.0	29144643.9		
Total (Cor.)	169.0	115226931.0	681816.2		

b. $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$ (This set of predictors does not contribute to explaining any of the variability in FVC.) We reject the null hypothesis that these predictors, as a set,

do not significantly predict FVC, as the set of predictors significantly predict FVC, $F(9,160) = 13.90$, $p < 0.0001$.

c. $R_C^2 = \frac{CSS(Regression)}{CSS(Total)} = \frac{50570942}{115226931} = 0.4389$

d. For all of the following calculations, the 2 individuals with any missing data were deleted before running the model.

i. Comparing intercept-only model to full model

$$Full\ Model: FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \beta_9(humid) + \varepsilon_i$$

$$Intercept - Only\ Model: FVC_i = \beta_0 + \varepsilon_i$$

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$$

$$F_{obs} = \frac{\frac{SSE(reduced) - SSE(full)}{dfE(reduced) - dfE(full)}}{SSE(full)/dfE(full)} = \frac{\frac{115226931 - 64655989}{169 - 160}}{64655989/160} = \frac{50570942/9}{404100} = 13.90$$

We reject the null hypothesis that these predictors, as a set, do not significantly predict FVC, as the set of predictors significantly predict FVC, $F(9,160) = 13.90$, $p < 0.0001$.

ii. Comparing intercept-only model to model with height as predictor

$$Larger\ Model: FVC_i = \beta_0 + \beta_1(height) + \varepsilon_i$$

$$Intercept - Only\ Model: FVC_i = \beta_0 + \varepsilon_i$$

$$H_0: \beta_1 = 0 \text{ in } FVC_i = \beta_0 + \beta_1(height) + \varepsilon_i$$

$$F_{obs} = \frac{\frac{SSE(smaller) - SSE(larger)}{dfE(smaller) - dfE(larger)}}{SSE(full)/dfE(full)} = \frac{\frac{115226931 - 76194753}{169 - 168}}{64655989/160} = \frac{39032178}{404100} = 96.59$$

Height provides information about FVC in a simple regression, F(1,160)=96.59, p<0.001. We reject the null hypothesis that height provides no information in predicting FVC, height is significantly related to FVC.

iii. Comparing model with weight and BMI as predictors to model with height, weight, and BMI as predictors

$$Larger\ Model: FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \varepsilon_i$$

$$Smaller\ Model: FVC_i = \beta_0 + \beta_2(weight) + \beta_3(BMI) + \varepsilon_i$$

$$H_0: \beta_1 = 0 \text{ in } FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \varepsilon_i$$

$$F_{obs} = \frac{\frac{SSE(smaller) - SSE(larger)}{dfE(smaller) - dfE(larger)}}{SSE(full)/dfE(full)} = \frac{\frac{70003745 - 69983719}{167 - 166}}{64655989/160} = \frac{20026}{404100} = 0.0496$$

After adjusting for weight and BMI, height provides no additional information in predicting FVC, F(1,160)=0.0496, p=0.82. We fail to reject the null hypothesis that height provides information about FVC after controlling for weight and BMI.

iv. Comparing full model to full model minus height

$$Full\ Model: FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \beta_9(humid) + \varepsilon_i$$

$$Reduced\ Model: FVC_i = \beta_0 + \beta_2(weight) + \beta_3(BMI) + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \beta_9(humid) + \varepsilon_i$$

$H_0: \beta_1 = 0$ in
$FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \beta_9(humid) + \varepsilon_i$

$$F_{obs} = \frac{\frac{SSE(reduced) - SSE(full)}{dfE(reduced) - dfE(full)}}{SSE(full)/dfE(full)} = \frac{\frac{64719380 - 64655989}{161 - 160}}{64655989/160} = \frac{63391}{404100} = 0.1569$$

After adjusting for all the other predictors in the model, height provides no additional information in predicting FVC, F(1,160)=0.1569, p=0.69. We fail to reject the null hypothesis that height provides information about FVC after controlling for the other variables in the model.

v. Comparing intercept-only model to model with model with height, weight, and BMI as predictors

$$Larger\ Model: FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \varepsilon_i$$

$$Intercept - Only\ Model: FVC_i = \beta_0 + \varepsilon_i$$

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$ in $FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \varepsilon_i$

$$F_{obs} = \frac{\frac{SSE(smaller) - SSE(larger)}{dfE(smaller) - dfE(larger)}}{SSE(full)/dfE(full)} = \frac{\frac{115226931 - 69983719}{169 - 166}}{64655989/160} = \frac{15081070}{404100} = 37.32$$

Height, weight, and BMI together provide additional information about FVC compared to a model for only the mean level of FVC, F(3,160)=37.32, p<0.001. We reject the null hypothesis that the body size variables provide no more information than the mean-only model. As a set, these three variables provide significant information about FVC.

vi. Comparing model with age, elevation, speed, temp, barometric pressure, and humidity with model with age, elevation, speed, temp, barometric pressure, humidity, height, weight, and BMI

$$Full\ Model: FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \beta_9(humid) + \varepsilon_i$$

$$Reduced\ Model: FVC_i = \beta_0 + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \beta_9(humid) + \varepsilon_i$$

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$ in $FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \beta_9(humid) + \varepsilon_i$

$$F_{obs} = \frac{\frac{SSE(reduced) - SSE(full)}{dfE(reduced) - dfE(full)}}{SSE(full)/dfE(full)} = \frac{\frac{99234383 - 64655989}{163 - 160}}{64655989/160} = \frac{11526131}{404100} = 28.523$$

As a set, height, weight, and BMI together provide additional information about FVC after controlling for all of the other variables in the model, F(3,160)=28.523, p<0.001. We reject the null hypothesis that the body size variables provide no more information after controlling for all of the other variables in the model. As a set, these three variables provide significant information about FVC after controlling for the other variables.

e. *Full Model*: $FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \beta_9(humid) + \varepsilon_i$

Reduced Model: $FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \varepsilon_i$

$H_0: \beta_9 = 0$ in $FVC_i = \beta_0 + \beta_1(height) + \beta_2(weight) + \beta_3(BMI) + \beta_4(age) + \beta_5(avg\ treadmill\ elevation) + \beta_6(avg\ treadmill\ speed) + \beta_7(temp) + \beta_8(barom) + \beta_9(humid) + \varepsilon_i$

$$F_{obs} = \frac{\frac{SSE(reduced) - SSE(full)}{dfE(reduced) - dfE(full)}}{SSE(full)/dfE(full)} = \frac{\frac{64768642 - 64655989}{161 - 160}}{64655989/160} = \frac{112653}{404100} = 0.2788, p = 0.5982$$

We fail to reject the null hypothesis that humidity has no effect on FVC after controlling for the other variables in the model, F(1,160)=0.2788, p=0.5982. After controlling for the other variables in the model, humidity does not significantly relate to FVC.

f. The body size variables together are significantly related to FVC, F(3,160)=37.32, p<0.001. They are significantly related even after controlling for all other variables in the model, F(3,160)=28.523, p<0.001. However, the variables are correlated with each other, as height is significantly related to FVC, F(1,160)=96.59, p<0.001, but not after controlling for weight and BMI, F(1,160)=0.0496, p=0.82.

g. Examining the added-in-order test, we see that height and weight are both highly related to FVC, but controlling for the other variables, neither height nor weight are significantly related (collinearity with BMI). The added-in-order also reveals that average treadmill elevation and average treadmill speed are significantly related to FVC as well. Again, controlling for the other variables in the model, neither variable individually is significantly related to FVC. Age is significantly related to FVC when controlling for the other variables, but is not significantly related to FVC when controlling for only height, weight, and BMI.

Examining parameter estimates from the added-in-order tests, we find that increasing height as well as increasing weight is associated with an increased FVC. Controlling for height, weight, BMI, and age, increasing average treadmill elevation as well as increasing treadmill speed are associated with increased FVC. While age is not significantly related to FVC controlling for height, weight, and BMI, an increased age is associated with increased FVC controlling for all other variables in the model.

3. This problem involves the FEV data described above. We will consider the model $FVC_i = \beta_0 + \beta_1 X_H + \beta_2 X_W + \beta_3 X_{BMI} + \beta_4 X_{Area} + \beta_5 X_{Age} + \beta_6 X_{avtrel} + \beta_7 X_{avtrsp} + \beta_8 X_{avtrel*avtrsp} + \beta_9 X_{temp} + \beta_{10} X_{barm} + \beta_{11} X_{hum} + \epsilon$.

a. Compute the following correlations, giving the interpretation of each, between FVC and age, and report tests of the hypotheses that each correlation equals zero.

i. The correlation between age and FVC, controlling both for all the other variables in the model

- **Hypotheses**:
  - $H_0: \rho_{(age,FVC|other\ variables)} = 0$
  - $H_A: \rho_{(age,FVC|other\ variables)} \neq 0$

- **Test Statistic**: $F_{obs} = \frac{\frac{[SSE(\text{no age}) - SSE(full)]}{[df_e(\text{no age}) - df_e(full)]}}{\frac{SSE(full)}{df_e(full)}} = \frac{\frac{[64828716.44 - 62761458.29]}{[159-158]}}{\frac{62761458.29}{158}} = 5.20426$
- **Degrees of Freedom**: $df_e(\text{no age}) = 159, df_e(full) = 158$
- **P-value**: $\Pr(F_{obs} > F_{(1,158)}) = 0.02387$
- **Decision**: Reject the null hypothesis
- **Interpretation**: There is a nonzero correlation between age and FVC after adjusting both for the other variables in the model.
- **Correlation**:
  - According to SAS, $\rho_{(age,FVC|other\ variables)} = 0.17857$
  - This suggests that, after controlling both variables for all of the other variables, with one increase in standard deviation of age, FVC is expected to increase by 0.17857 standard deviations.

ii. The correlation between age and FVC, controlling only age for all the other variables in the model

- **Hypotheses**:
  - $H_0: \rho_{FVC(age|other\ variables)} = 0$
  - $H_A: \rho_{FVC(age|other\ variables)} \neq 0$
- **Test Statistic**:
  - First, used SAS to model obtain studentized residuals of age=(other variables)
  - Second, used SAS to model avfvc=(studentized residuals)
    - $F_{obs} = \frac{\frac{[SSE(\beta_0) - SSE(full)]}{[df_e(\beta_0) - df_e(full)]}}{\frac{SSE(full)}{df_e(full)}} = \frac{\frac{[CSS(model)]}{[169-168]}}{\frac{SSE(full)}{df_e(full)}} = \frac{\frac{[2067258.2]}{[169-168]}}{\frac{113159672.8}{168}} = 3.07$
- **Degrees of Freedom**: $df_e(\beta_0) = 169, df_e(full) = 168$
- **P-value**: $\Pr(F_{obs} > F_{(1,168)}) = 0.0816$
- **Decision**: Fail to reject the null hypothesis
- **Interpretation**: After controlling the other variables on age, there isn't enough evidence to suggest a significant correlation between age (adjusted) and FVC.
- **Correlation**: $r_{FVC(age|other\ variables)} = r(FVC_i, \hat{e}_{age}) = 0.13394$. If this correlation was statistically significant, it would suggest that with one increase in standard deviation of age (adjusted), FVC (unadjusted) is expected to increase by 0.13570 standard deviations.

iii. The simple correlation between age and FVC (not controlling for any other variables)

- **Hypotheses**: $H_0: \rho_{FVC,age} = 0$ vs. $H_A: \rho_{FVC,age} \neq 0$
- **Test Statistic**:
  - $F_{obs} = \frac{\frac{[SSE(\beta_0) - SSE(\beta_0,age)]}{[df_e(\beta_0) - df_e(\beta_0,age)]}}{\frac{SSE(\beta_0,age)}{df_e(\beta_0,age)}} = \frac{\frac{[115226930.97 - 112547661.64]}{[169-168]}}{\frac{112547661.64}{168}} = 3.99935$
- **Degrees of Freedom**: $df_e(\beta_0) = 169, df_e(full) = 168$
- **P-value**: $\Pr(F_{obs} > F_{(1,168)}) = 0.047129$
- **Decision**: Reject the null hypothesis
- **Interpretation**: There is enough evidence to suggest a simple correlation between age and FVC.
- **Correlation**: $r_{FVC,age} = 0.15249$. This suggests that with one increase in standard deviation of age (unadjusted), FVC (unadjusted) is expected to increase by 0.15249 standard deviations.

b. Provide and interpret the following diagnostics (include subject ID when appropriate) for the regression model.

(b)

i.

the largest 5 studentized residuals in absolute values are:

Subject ID	Studentized residuals in absolute value
60	3.882
185	2.904
49	2.903
181	2.660
99	2.267

ii.

To test whether the studentized residuals are normal, we can use one of:

1) Shapiro-Wilk Test;

2) Kolmogorov-Smirnov Test;

3) Cramer-von Mises Test;

4) Anderson-Darling Test.

The null hypothesis is that the studentized residuals follow normal distribution, and the followings are the corresponding test statistics and the p-values:

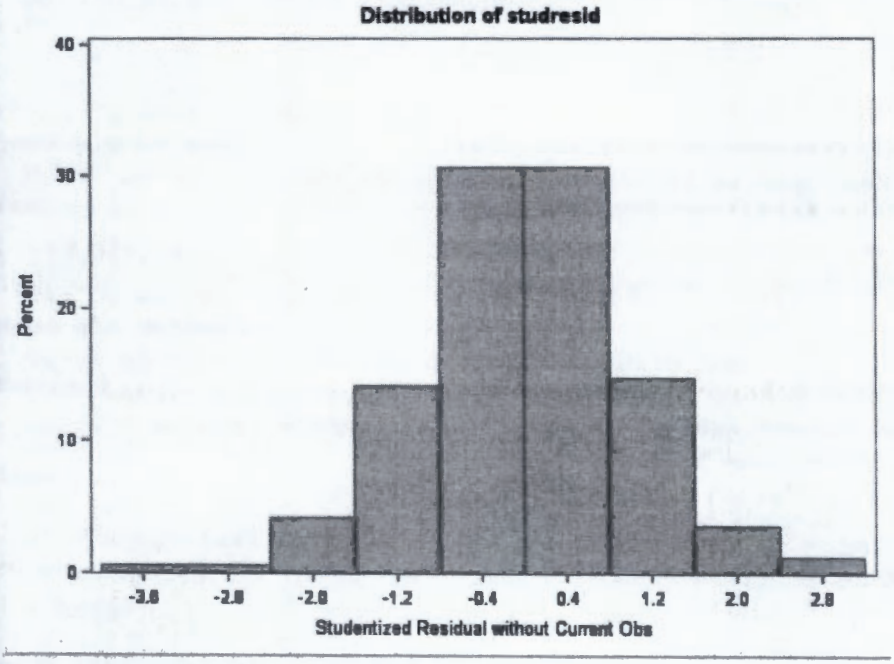
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.989513	Pr < W	0.2420
Kolmogorov-Smirnov	D	0.044551	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.039587	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.308589	Pr > A-Sq	>0.2500

All the tests show p-values that are greater than 0.05, so that we fail to reject the null hypothesis. The result can be interpreted as the following: assuming that the studentized residuals are normally distributed, then the probability of observing the data that are as extreme as ours or more extreme is larger than 0.05, which is not too small for us to question the null hypothesis.

Note that the type of test should be decided before conducting the test to avoid "p shopping", although the test statistics and p-values are all reported in the table above.

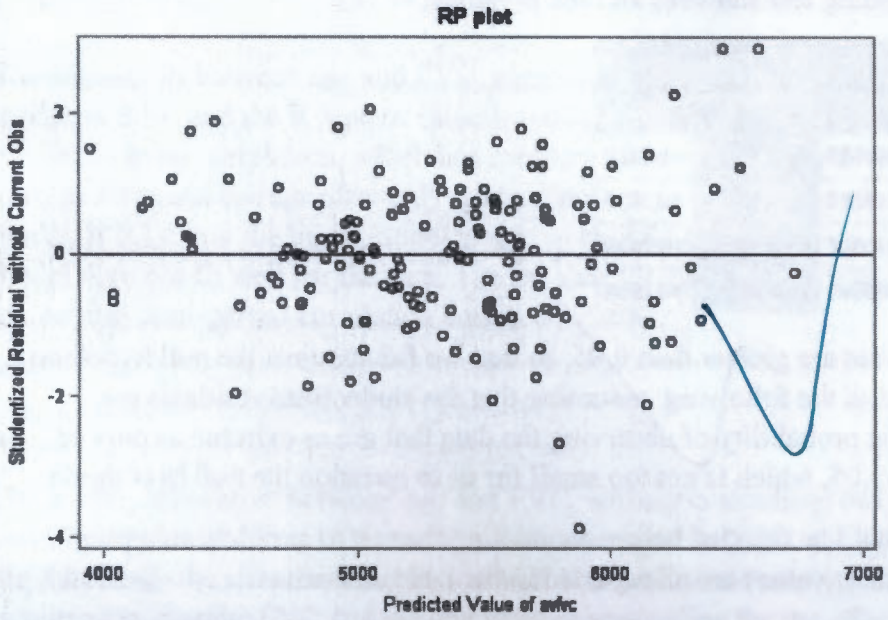
iii.

The histogram of the studentized residuals is plotted as the following:



The histogram can be interpreted as the following: the shape of the studentized residuals appears to be normal in this histogram, which coincide with our hypothesis testing in part ii.

iv. the RP plot of studentized residuals is plotted as the following:



The RP plot can be interpreted as the following: from the RP plot we can see that the linearity assumption, the homogeneity of variance assumption, and Gauss error assumption holds, because we cannot see any nonlinear pattern or heterogeneity of variance of the residuals. Also the studentized residuals appear to be normal as well.

BIOS663 Homework 4
Due Monday, April 8 in class.

1. The following questions are on the data and model described in Q3 of HW3:

   - Examine the tolerances and variance inflation factors from this model. Do you think any collinearity is present based on the tolerance and VIF? Why or why not?
   - Conduct an eigenanalysis of the scaled SSCP and correlation matrices, presenting a table formatted like Table 8.6.2 in Muller and Fetterman.
     (a) Does there appear to be any collinearity between the intercept and the covariates? Why or why not? If so, list the variables?
     (b) Does there appear to be any collinearity among the covariates? Why or why not? If so, list the variables?

2. Find the Box-Cox transformation of the simulated data (BoxCox.dat) and compare the residual plots of the raw and transformed data.

3. Investigators are interested in the effect of dermal nicotine exposure in a population of Latino tobacco workers in North Carolina. (Nicotine can be absorbed from tobacco leaves through the skin and can cause nicotine poisoning, which is characterized by nausea, vomiting, headache, and dizziness.) Data were collected on tobacco work tasks and risk factors for exposure to nicotine during a summer tobacco work season. Nicotine exposure was measured by levels of cotinine, a nicotine metabolite, contained in saliva. Other covariates of interest include age, body mass index, education, work conditions (working in wet conditions is believed to increase nicotine absorption), type of tobacco work ("priming" refers to picking or harvesting the tobacco and is expected to result in highest nicotine exposures, "barning" refers to putting the harvested tobacco into a barn for curing, "topping" refers to breaking the flower off the top of the plant, and "other" refers to farm work that does not involve tobacco contact, such as driving a truck), and smoking (smokers would also have nicotine exposure through cigarettes, and it is not known whether exposure to tobacco leaves would increase cotinine levels to a similar extent in both smokers and non-smokers).

   The variables are available in the file tobacco.dat and listed in the following order.

   - COTININE: salivary cotinine concentration (in ng/mL)
   - AGE: age (in years)
   - BMI: body mass index (in kg/m^2)
   - EDUC: years of education
   - WET: takes value 1 if work conditions on day of measurement were wet and takes value 0 otherwise

- TASK: takes value 1 for priming, 2 for barning, 3 for topping, and 4 for other work not involving tobacco contact
- LNNSMOKE: natural logarithm of (1 + number of cigarettes smoked per day)

To report a test, provide H_0, the test statistic, the degrees of freedom, the p-value, the decision (accept/reject H_0), and an interpretation of the result in terms of the subject matter.

- One-Way ANOVA: For these questions, use the log of salivary cotinine as the response and task as the only predictor.
  - Report a test of whether all cell means are equal.
  - If your overall test of the task effect was significant, examine all pairwise comparisons using the Scheffe correction. Summarize your findings in a table including columns for the estimated mean difference, degrees of freedom, F statistic, p-value, and a Scheffe confidence interval for the mean difference. Explain your findings in language the investigator can understand.
  - Provide a table of parameter estimates and standard errors using (a) cell mean coding and (b) reference cell coding, and give the interpretations of parameters in both coding schemes. In addition, provide the $\mathbf{C}$ and $\boldsymbol{\theta}_0$ matrices used to test the hypothesis that average cotinine levels for workers involved in priming are greater than the average continine levels for all other workers.
- Two-Way ANOVA: For these questions, use the log of salivary cotinine as the response and task and wet as predictors.
  - Fit the two-way ANOVA model with full interaction, and interpret all parameter estimates in your model, clearly stating which coding scheme you used. Discuss the validity of the HILE Gauss assumptions for this model.
  - *Based on this model*, create a table of the estimated mean log cotinine levels, associated standard errors, and how each estimated mean is obtained from the model parameters (e.g., $\widehat{\beta}_0 + \widehat{\beta}_1$) for each task-wet combination.
- The Full Model in Every Cell: For these questions, use the log of salivary cotinine as the response and task, and lnnsmoke as predictors.
  - Fit the full model in every cell. Provide and interpret estimates of all parameters for this model.
  - Report an appropriate test of whether task is related to cotinine levels. If this test is significant, report step-down tests to determine exactly where differences lie. For all tests reported, be sure to state H_0 clearly and give explicit justification for which tests were used and why.

BIOS 663 Homework 4
4/8/2019
Problem 1

The following questions are on the data and model described in Q3 of HW3.

part i:

Examine the tolerances and variance inflation factors from this model. Do you think any collinearity is present based on tolerance and VIF? Why or why not?Define tolerance as follows:

$$T_j = 1 - R_j^2$$

where $R_j^2 = R^2(X_j, \{X_1, \ldots X_{j-1}, X_{j+1}, \ldots X_{p-1}\})$ is the squared multiple correlation. Tolerances close to 1 are good, where tolerances close to 0 show worse collinearity.

Define the variance inflation factor as follows:

$$VIF_j = \frac{1}{1 - R_j^2} = \frac{1}{T_j}$$

A VIF close to 1 is good, where a VIF implies worse collinearity as it approaches infinity. The R-Code below calculates the Tolerance and VIF values for the model.

```
fit = lm(AVFVC ~ HEIGHT + WEIGHT + BMI + AREA + AGE + AVTREL + AVTRSP + AVTREL*AVTRSP + TEMP + BAROM + HUMID, data = dat2)
VIF = vif(fit)
Tol = 1/VIF
df = (rbind(VIF, Tol))
df %>% knitr::kable(align = c("c", "c"))
```

	HEIGHT	WEIGHT	BMI	AREA	AGE	AVTREL	AVTRSP	TEMP	BAROM	HUMID	AVTREL:AVTRSP
VIF	458.0476405	703.4462861	177.4503720	1364.8975242	1.0833448	580.389689	78.3930237	29.0137857	1.0590953	29.1669187	795.4489506
Tol	0.0021832	0.0014216	0.0056354	0.0007327	0.9230672	0.001723	0.0127562	0.0344664	0.9442021	0.0342854	0.0012572

Based on these values, it appears that there is a lot of collinearity present. This is because very few VIF values are close to 1, and many of the tolerance values are close to 0.

part ii:

Conduct an eigenanalysis of the scaled SSCP and correlation matrices, presenting a table formatted like Table 8.6.2 in Muller and Fetterman.

The Scaled SSCP Matrix can be defined as follows:

$$SSCP_s = D_s^{-0.5}(X'X)D_s^{-0.5}$$

where X is the design matrix includign the intercept, and D_s is a diagonal matrix with elements extracted from the diagonal of $X'X$.

and the correlation matrix as follows:

$$R = D_c^{-0.5} C D_c^{-0.5}$$

where C is the covariance matrix of the centered design matrix excluding the intercept, and D_c is a diagonal matrix of the extracted diagonal values of C.

The following R code calculates these two matrices for the model above, and performs an eigenanalysis on them both.

```
cov_int <- dat2 %>% mutate(INT = 1, AVTRELTRSP = AVTREL*AVTRSP) %>% select(INT, HEIGHT, WEIGHT, BMI, AREA, AGE, AVTREL, AVTRSP, AVTRELTRSP, T
EMP, BAROM, HUMID) %>% as.matrix()
xtx = t(cov_int)%*%cov_int
Ds_half <- diag(diag(xtx)^-0.5)
sscp <- Ds_half %*% xtx %*% Ds_half
eig_sscp <- eigen(sscp)$values
PCs_sscp <- prcomp(sscp)[2]
CI_sscp <- sqrt(eig_sscp[1]/eig_sscp)

covariates <- dat2 %>% mutate(AVTRELTRSP = AVTREL*AVTRSP) %>% select(HEIGHT, WEIGHT, BMI, AREA, AGE, AVTREL, AVTRSP, AVTRELTRSP, TEMP, BAROM,
HUMID)
cov_center <- apply(covariates, 2, function(y) y - mean(y))
C <- (t(cov_center)%*%cov_center)/dim(cov_center)[1]
Dc_half <- diag(diag(C)^-0.5)
R <- Dc_half %*% C %*% Dc_half
eig_corr <- eigen(R)$values
CI_corr <- sqrt(eig_corr[1]/eig_corr)
PCs_corr <- prcomp(R)[2]

df <- data.frame("Eigenvalue" = c("Correlation Matrix", eig_corr), "Condition Index" = c("Correlation Matrix", CI_corr))
df2 <- data.frame("Eigenvalue" = c("Scaled SSCP", eig_sscp), "Condition Index" = c("Scaled SSCP", CI_sscp))

df %>% knitr::kable(align = c("c", "c"))
```

Eigenvalue	Condition.Index
Correlation Matrix	Correlation Matrix
3.00984215183995	1
2.44782688677429	1.10887223583127
2.02476480717731	1.21922699035498
1.11320126901436	1.64431498130338
1.01325012874562	1.72350888345776
0.809279431894358	1.92851316812643
0.561095110548582	2.31608032840747
0.0177075849003461	13.0374361244415
0.00187373597726271	40.0790724581539
0.000705172710081123	65.3317229298841
0.000453720417828643	81.4474887486485

```
df2 %>% knitr::kable(align = c("c", "c"))
```

Eigenvalue	Condition.Index
Scaled SSCP	Scaled SSCP
11.9049479582918	1
0.0360382910658672	18.1753028593325
0.0293708868488598	20.132848271369
0.0161788245373072	27.1262817394066
0.00669911241875652	42.1555847347362
0.0049048528907778	49.2663919299516
0.0015549706468733	87.4989120284385
0.000251548277639727	217.546988935216
3.7467685203547e-05	563.683495875739
9.67075357565232e-06	1109.51605795048
4.86755151319751e-06	1563.89817359537

1.54903175502382e-06 2772.25764760603

(a):

Does there appear to be any collinearity between the intercept and the covariates? Why or why not? If so, list the variables.

Since the eigenvalues from the Scaled SSCP (which includes the intercept) show several eigenvalues near 0 and condition indices above 30 (namely the last 8), we know that there does appear to be collinearity issues. To identify which covariates this collinearity is between, we take a look at the last few PCs below.

```
PCs_sscp$rotation[,9:12]
```

```
##                PC9          PC10          PC11          PC12
##  [1,] -0.362557672  0.2106087110  0.6871965147  0.1473741435
##  [2,] -0.302287651 -0.3282233486 -0.4589345265  0.4834046812
##  [3,]  0.228929531  0.4172069832  0.1173983196  0.3239142829
##  [4,] -0.186467777 -0.2773877633 -0.1708233087 -0.0167347071
##  [5,] -0.097183258 -0.3259677478  0.1402711189 -0.7151909522
##  [6,]  0.008240427  0.0007674032  0.0006411418 -0.0004944857
##  [7,] -0.093561530  0.3902109785 -0.2888650047 -0.2038318770
##  [8,] -0.094260468  0.4147408015 -0.2908961834 -0.2059574809
##  [9,]  0.100295075 -0.3935164095  0.2925559731  0.2064610370
## [10,] -0.025805362  0.0012700485 -0.0127654815 -0.0270116541
## [11,]  0.806990526 -0.1072759346 -0.0237080809 -0.0096174923
## [12,]  0.029751201 -0.0035722176  0.0110939585  0.0220843112
```

From the PCA analysis, we can see that the covariates with the largest departures from 0 in the last four PCs are covariates 1 (i.e. the intercept), 2 (height), 3 (weight), 5 (area), 7 (avtrel), 8 (avtrsp), and 9 (avtrel*avtrsp).

(b):

Does there appear to be any collinearity among the covariates? Why or why not? If so, list the variables.

Since the eigenvalues from the Correlation Matrix (which does NOT includes the intercept) show several eigenvalues near 0 and condition indices above 30 (namely the last three), we know that there does appear to be collinearity issues. To identify which covariates this collinearity is between, we take a look at the last few PCs below.

```
PCs_corr$rotation[,10:11]
```

```
##               PC10         PC11
##  [1,]  0.108339574 -0.507595144
##  [2,]  0.544305474 -0.014314017
##  [3,] -0.130059319 -0.310181321
##  [4,] -0.544577040  0.582884584
##  [5,] -0.004893314 -0.001127961
##  [6,] -0.390534996 -0.350817108
##  [7,] -0.138224165 -0.124569696
##  [8,]  0.454476015  0.409062361
##  [9,] -0.012113720  0.009805947
## [10,] -0.002270905 -0.002276515
## [11,]  0.012911758 -0.012268259
```

From the PCA analysis, we can see that the covariates with the largest departures from 0 in the last four PCs are covariates 1 (height), 3 (BMI), 4 (area), 6 (avtrel), and 8 (avtrel*avtrsp).

Problem 2

Find the Box-Cox Transformation of the simulated data (BoxCox.dat) and compare the residual plots of the raw and transformed data.

The Box-Cox Transformations are a family of transformations of the response varaibles defined as:

$$Y_i(\pi) = \{ \frac{Y_i^{\pi} - 1}{\pi Y^{*(\pi-1)}} \quad \pi \neq 0 \quad \& \quad Y^{*} ln(Y_i) \quad \pi = 0 \}$$

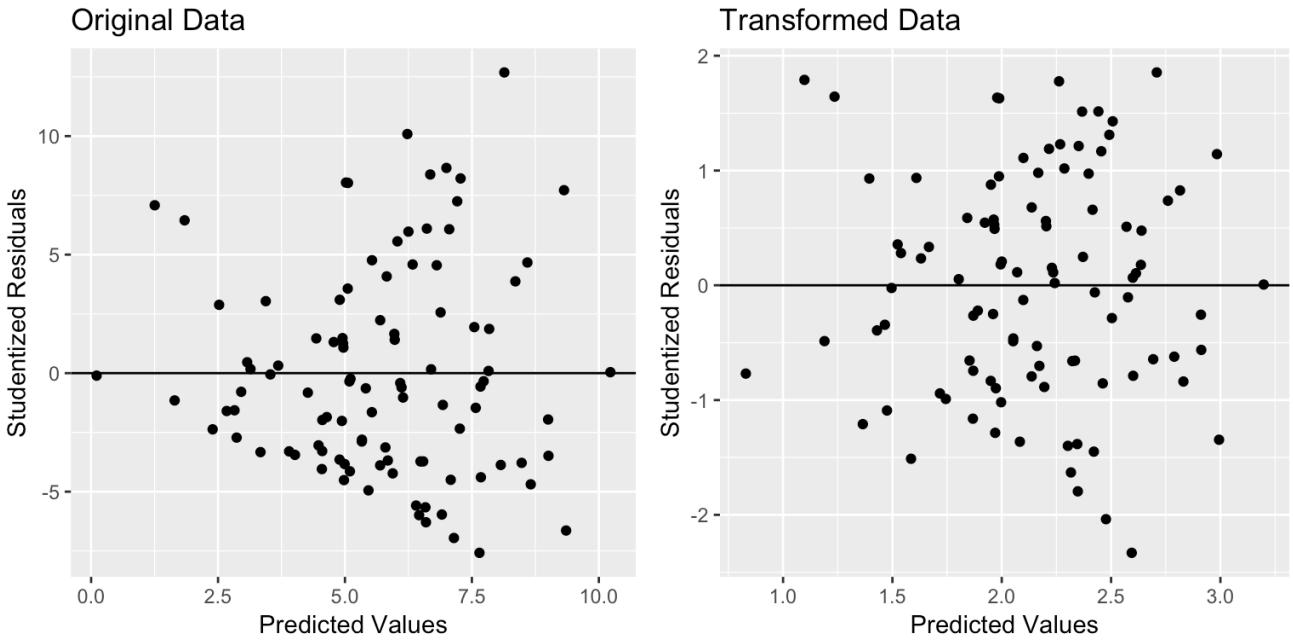
where $Y^* = (\Pi_{i=1}^N Y_i)^{1/N}$. This corresponds to a transformation that is y^{π} for $\pi \neq 0$ and $log(y)$ otherwise. The transformation above puts the SSE of these on the same scale for the purpose of comparison and choosing the best π. We try the values of π between -1 and 1 incremented every 0.25 to compare the likelihoods and find the best transformation.

```
fit_bc <- lm(bc$V2 ~ bc$V1)
# Transformation Analysis
cols <- MASS::boxcox(fit_bc, seq(-1,1,1/4), plotit = FALSE)$x
like <- MASS::boxcox(fit_bc, seq(-1,1,1/4), plotit = FALSE)$y %>% as.matrix() %>% t()
knitr::kable(like, col.names = cols)
```

-1	-0.75	-0.5	-0.25	0	0.25	0.5	0.75	1
-685.0735	-549.0958	-425.2876	-325.0825	-264.4143	-241.4822	-239.6977	-248.8666	-264.7483

From the values above, we can see that the choice of π with the smallest likelihood is $\pi = 0.5$. Below, we transform the data using this result, and compare the residual plots of the two datasets.

```
lambda = 0.5
bc$V3 <- (bc$V2^lambda)
fit_bc_2 <- lm(bc$V3 ~ bc$V1)

plot1 <- ggplot() + geom_point(aes(fit_bc$fitted.values, fit_bc$residuals)) + geom_hline(aes(yintercept = 0)) + labs(x = "Predicted Values",
y = "Studentized Residuals", title = "Original Data")
plot2 <- ggplot() + geom_point(aes(fit_bc_2$fitted.values, fit_bc_2$residuals)) + geom_hline(aes(yintercept = 0)) + labs(x = "Predicted Value
s", y = "Studentized Residuals", title = "Transformed Data")
cowplot::plot_grid(plot1, plot2, nrow = 1)
```



From the plots above, it is clear that the transformation of the data yields better assumption validations than the original data. In particular, the graph on the left of the original data seems to fan out (i.e. more extreme residuals) as the predicted values are increased. The residuals are more randomly distributed around the x-axis in the transformed data.

Problem 3

part i: One-Way ANOVA:

For these questions, use the log of salivary cotinine as the response and task as the only predictor.

```
tobacco1 <- tobacco %>% mutate(LOGCOT = log(COTININE),
                               TASK1 = case_when(TASK == 1 ~ 1, TASK != 1 ~ 0),
                               TASK2 = case_when(TASK == 2 ~ 1, TASK != 2 ~ 0),
                               TASK3 = case_when(TASK == 3 ~ 1, TASK != 3 ~ 0),
                               TASK4 = case_when(TASK == 4 ~ 1, TASK != 4 ~ 0))
```

(a):

Report a test of whether all cell means are equal.

Consider the following model using the cell mean coding scheme:

$$y = \beta_1 I_{T1} + \beta_2 I_{T2} + \beta_3 I_{T3} + \beta_4 I_{T4}$$

where y is the log cotinine, and I_{Ti} is the indicator function associated with the ith task. In order to test whether all cell means are equal, we want to test the following set of hypotheses:

$$H_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4$$

which is equivalent to:

$$H_0 = \beta_1 - \beta_2 = 0, \beta_1 - \beta_3 = 0, \beta_1 - \beta_4 = 0$$

In order to test these hypotheses, we can use the overall F test, where:

$$F = (SSH/a)/\hat{\sigma}^2 \sim F_{G-1,n-G}$$

where $SSH = (\hat{\theta} - \theta_0)' M^{-1} (\hat{\theta} - \theta_0)$, $G = 4, n = 694$, and $^2 = $. It should also be noted that $M = C(X'X)^{-1}C'$.

```
X = tobacco1 %>% select(TASK1, TASK2, TASK3, TASK4) %>% as.matrix()
fit = lm(LOGCOT ~ -1 + TASK1 + TASK2 + TASK3 + TASK4, data = tobacco1)
thetahat = c((fit %>% summary)$coefficients[1,1] - (fit %>% summary)$coefficients[2,1],
             (fit %>% summary)$coefficients[1,1] - (fit %>% summary)$coefficients[3,1],
             (fit %>% summary)$coefficients[1,1] - (fit %>% summary)$coefficients[4,1])
mse = sum(fit$residuals^2)/(694-4)
C = matrix(c(1, 1, 1, -1, 0, 0, 0, -1, 0, 0, 0, -1), nrow = 3)
M = C %*% solve(t(X) %*% X) %*% t(C)
ssh = t(thetahat) %*% solve(M) %*% thetahat
f_obs = (ssh/3)/mse
p = 1-pf(f_obs, 4-1, 694-4)
## CAN ALSO USE linearHypothesis(fit, C)
```

From the code above, the test statistic $F = 116.2032527$ and the p-value is approximately 0. This means that we can reject the null hypothesis that all four cell means are equal. In other words, there is evidence to reject the fact that the four types of tobacco work have the same mean log cotinine level.

(b):

If your overall test of the task effect was significant, examine all pairwise comparisons using the Scheffe correction. Summarize your findings in a table including columns for the estimated mean difference, degrees of freedom, F statistic, p-value, and a Scheffe confidence interval for the mean difference. Explain your findings in language the investigator can understand.

Since the overall test of the task effect in (a) was significant, we will go forward with the pairwise comparisons. Since TASK has four levels, there will be $4 * (4 - 1)/2 = 6$ pairwise comparisons. Scheffe's correction provides a general technique for account for the fact that we are performing 6 tests.

To find the F statistic, we can take the square of the t statistic as follows:

$$F = t^2 = (\frac{\hat{\beta_i} - \hat{\beta_j}}{\sqrt{MSE(1/n_i + 1/n_j)}})^2 \sim F_{G-1,n-G}$$

where β_i and n_i are the mean log cotinine level and sample size for the ith task level. MSE is the mean squared error as calculated in the previous test. The critical region for this F test can be calculate by multiplying the F statistic by $G - 1 = 4 - 1 = 3$ to account for multiplicity in testing.

```
scheffe <- ScheffeTest(aov(LOGCOT ~ factor(TASK), data = tobacco1))
f1 <- ((summary(fit)$coefficients[1,1] - summary(fit)$coefficients[2,1]) /sqrt(mse*(1/sum(tobacco1$TASK1) + 1/sum(tobacco1$TASK2))))^2
f2 <- ((summary(fit)$coefficients[1,1] - summary(fit)$coefficients[3,1]) /sqrt(mse*(1/sum(tobacco1$TASK1) + 1/sum(tobacco1$TASK3))))^2
f3 <- ((summary(fit)$coefficients[1,1] - summary(fit)$coefficients[4,1]) /sqrt(mse*(1/sum(tobacco1$TASK1) + 1/sum(tobacco1$TASK4))))^2
f4 <- ((summary(fit)$coefficients[2,1] - summary(fit)$coefficients[3,1]) /sqrt(mse*(1/sum(tobacco1$TASK2) + 1/sum(tobacco1$TASK3))))^2
f5 <- ((summary(fit)$coefficients[2,1] - summary(fit)$coefficients[4,1]) /sqrt(mse*(1/sum(tobacco1$TASK2) + 1/sum(tobacco1$TASK4))))^2
f6 <- ((summary(fit)$coefficients[3,1] - summary(fit)$coefficients[4,1]) /sqrt(mse*(1/sum(tobacco1$TASK3) + 1/sum(tobacco1$TASK4))))^2
df <- data.frame(Diff = scheffe$`factor(TASK)`[,1], DF = "(3,690)", f = c(f1, f2, f3, f4, f5, f6), pval = scheffe$`factor(TASK)`[,4], CI_L =
scheffe$`factor(TASK)`[,2], CI_U = scheffe$`factor(TASK)`[,3])
df %>% knitr::kable(align = c("c", "c"))
```

	Diff	DF	f	pval	CI_L	CI_U
2-1	-0.9207815	(3,690)	19.57598	0.0002350	-1.503991	-0.3375720
3-1	-1.6738481	(3,690)	131.87148	0.0000000	-2.082328	-1.2653684
4-1	-2.6992523	(3,690)	332.68364	0.0000000	-3.113975	-2.2845298
3-2	-0.7530666	(3,690)	12.98968	0.0049075	-1.338617	-0.1675167
4-2	-1.7784708	(3,690)	71.37797	0.0000000	-2.368393	-1.1885490
4-3	-1.0254042	(3,690)	47.25875	0.0000000	-1.443412	-0.6073970

From the table above, it appears that all pairwise null hypotheses can be rejected. This means there is evidence to suggest that every mean log cotinine level for a certain task level is different than the mean log cotinine level for any other task level.

(c):

Provide a table of parameter estimates and standard errors using (a) cell mean coding and (b) reference cell coding, and give the interpretations of pa- rameters in both coding schemes. In addition, provide the C and θ_0 matrices used to test the hypothesis that average cotinine levels for workers involved in priming are greater than the average continine levels for all other workers.

For the cell mean coding, we use the model proposed in (a).

```
summary(fit)$coefficients
```

```
##       Estimate Std. Error  t value      Pr(>|t|)
## TASK1 4.508557  0.1022201 44.10636 5.762099e-203
## TASK2 3.587775  0.1812765 19.79172  2.168761e-69
## TASK3 2.834709  0.1039098 27.28047 9.869409e-112
## TASK4 1.809304  0.1070123 16.90745  6.478023e-54
```

The parameter estimates and standard errors are given in the code summary above. The interpretations are as follows: $beta_1$ is the mean log cotinine level for priming, $beta_2$ is the mean log cotinine level for barning, $beta_3$ is the mean log cotinine level for topping, and $beta_4$ is the mean log cotinine level for work not involving tobacco contact.

For the reference cell coding, we consider the following model:

$$y = \beta_1 + \beta_2 I_{T2} + \beta_3 I_{T3} + \beta_4 I_{T4}$$

where y is the log cotinine, and I_{Ti} is the indicator function associated with the ith task. It should be noted that TASK1 is the reference.

```
fit_ref = lm(LOGCOT ~ TASK2 + TASK3 + TASK4, data = tobacco1)
summary(fit_ref)$coefficients
```

```
##               Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)  4.5085566  0.1022201  44.106361 5.762099e-203
## TASK2       -0.9207815  0.2081109  -4.424476  1.123234e-05
## TASK3       -1.6738481  0.1457607 -11.483531  4.699458e-28
## TASK4       -2.6992523  0.1479884 -18.239617  5.849323e-61
```

Again, the parameter estimates and standard errors are given in the code summary above. The interpretations are different than for the cell mean coding and are as follows: $beta_1$ is the intercept which again is the mean log cotinine level for priming, which is the reference level. $beta_2$ is the difference between the mean log cotinine level for barning and the mean log cotinine level for priming. Similarly, $beta_3$ is now the difference btween the mean log cotinine level for topping and the mean log cotinine level for priming, and $beta_4$ is the difference between the mean log cotinine level for work not involving tobacco contact and the mean log cotinine level for priming.

The TASK value corresponding with priming is 1, and so we want to test that the mean cotinine level for TASK1 is greater than the mean cotonine level for all others. We can test the following hypothesis using the cell mean coding:

$$H_0 = \beta_1 = (\beta_2 + \beta_3 + \beta_4)/3 \quad vs. \quad H_A = \beta_1 > (\beta_2 + \beta_3 + \beta_4)/3$$

This null hypothesis corresponds to:

$$H_0 = \beta_1 - \frac{1}{3}\beta_2 - \frac{1}{3}\beta_3 - \frac{1}{3}\beta_4 = 0$$

so $\theta_{\mathbf{0}} = [0]$ and $\mathbf{C} = [1 \quad -1/3 \quad -1/3 \quad -1/3]$

part ii: Two-Way ANOVA:

For these questions, use the log of salivary cotinine as the response and task and wet as predictors.

```
tobacco$WET <- tobacco$WET %>% as.factor()
tobacco$TASK <- tobacco$TASK %>% as.factor()
tobacco2 = tobacco %>% mutate(LOGCOT = log(COTININE),
                              WET0TASK1 = case_when(WET == 0 & TASK == 1 ~ 1,
                                                   WET != 0 | TASK != 1 ~ 0),
                              WET1TASK1 = case_when(WET == 1 & TASK == 1 ~ 1,
                                                   WET != 1 | TASK != 1 ~ 0),
                              WET0TASK2 = case_when(WET == 0 & TASK == 2 ~ 1,
                                                   WET != 0 | TASK != 2 ~ 0),
                              WET1TASK2 = case_when(WET == 1 & TASK == 2 ~ 1,
                                                   WET != 1 | TASK != 2 ~ 0),
                              WET0TASK3 = case_when(WET == 0 & TASK == 3 ~ 1,
                                                   WET != 0 | TASK != 3 ~ 0),
                              WET1TASK3 = case_when(WET == 1 & TASK == 3 ~ 1,
                                                   WET != 1 | TASK != 3 ~ 0),
                              WET0TASK4 = case_when(WET == 0 & TASK == 4 ~ 1,
                                                   WET != 0 | TASK != 4 ~ 0),
                              WET1TASK4 = case_when(WET == 1 & TASK == 4 ~ 1,
                                                   WET != 1 | TASK != 4 ~ 0))
```

(a):

Fit the two-way ANOVA model with full interaction, and interpret all parameter estimates in your model, clearly stating which coding scheme you used. Discuss the validity of the HILE Gauss assumptions for this model.

Consider the following model using the cell mean coding scheme:

$$y = \beta_1 I_{W0,T1} + \beta_2 I_{W1,T1} + \beta_3 I_{W0,T2} + \beta_4 I_{W1,T2} + \beta_5 I_{W0,T3} + \beta_6 I_{W1,T3} + \beta_7 I_{W0,T4} + \beta_8 I_{W1,T4}$$

where y is the log cotinine, and $I_{Ti,Wj}$ is the indicator function associated with the ith task and jth wet categories.

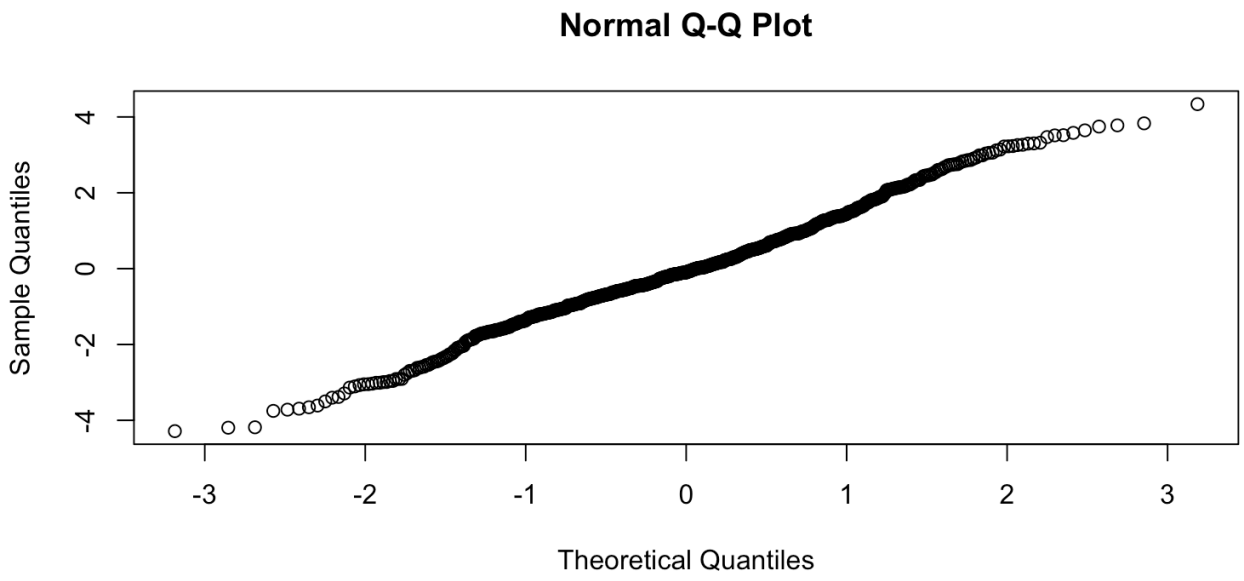
```
fit <- lm(LOGCOT ~ -1 + WET0TASK1 + WET1TASK1 + WET0TASK2 + WET1TASK2 + WET0TASK3 + WET1TASK3 + WET0TASK4 + WET1TASK4, data = tobacco2)
summary(fit)$coefficients
```

```
##           Estimate Std. Error   t value      Pr(>|t|)
## WET0TASK1 4.269337  0.1854712 23.018868  2.548339e-87
## WET1TASK1 4.613116  0.1226196 37.621351 6.459772e-169
## WET0TASK2 3.542748  0.2271549 15.596174  3.665240e-47
## WET1TASK2 3.667024  0.3013550 12.168449  5.528611e-31
## WET0TASK3 2.688185  0.2152536 12.488456  2.142659e-32
## WET1TASK3 2.879303  0.1187505 24.246652  2.808906e-94
## WET0TASK4 1.808889  0.1238562 14.604754  2.911782e-42
## WET1TASK4 1.810534  0.2130902  8.496563  1.211849e-16
```

The estimates for each of the parameters are shown in the summary statistics above. Since the cell mean coding is used, the interpretations are clear; each parameter represents the mean log cotinine level for the combination of WET and TASK levels listed. For example, β_1 is estimated by 4.269337 noted by the WET0TASK1 indicator variable.

In terms of HILE Gauss assumptions for this model, the only assumptions that are generally checked for ANOVA are H, I, and Gauss. The independence assumption is dependent on the design and the sampling scheme, and from the description of the design, I do not see any issues that would question the validity of the independence assumption. In terms of the homogeneity and gaussian errors assumptions, we can perform tests as done below to check these assumptions. We also note that the design is unbalanced (i.e. the sample size per cell ranges from 25 to 161), which is something to consider when using this model.

```
qqnorm(fit$residuals)
```



The linearity of the QQ-Plot above verifies the gaussian errors assumption.

```
leveneTest(LOGCOT ~ TASK*WET, data = tobacco2, center=median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group   7  18.635 < 2.2e-16 ***
##       686
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the Levene Test above, the p-value is < 2.2e-16, which means we reject the hypothesis that the homoegeneity of variance asssumption is satisfied. We should proceed with caution when using this model.

(b):

Based on this model, create a table of the estimated mean log cotinine levels, associated standard errors, and how each estimated mean is obtained from the model parameters (e.g., $\hat{\beta}_0 + \hat{\beta}_1$) for each task-wet combination.

```
beta <- intToUtf8(946)
params <- summary(fit)$coefficients[1:8,1:2] %>% as.data.frame()
params <- data.frame(params, "Parameter" = paste0(beta, c(1:8)))
params %>% knitr::kable(align = c("c", "c"))
```

	Estimate	Std..Error	Parameter
WET0TASK1	4.269337	0.1854712	β1
WET1TASK1	4.613116	0.1226196	β2
WET0TASK2	3.542748	0.2271549	β3
WET1TASK2	3.667024	0.3013550	β4
WET0TASK3	2.688185	0.2152536	β5
WET1TASK3	2.879303	0.1187505	β6
WET0TASK4	1.808889	0.1238562	β7
WET1TASK4	1.810534	0.2130902	β8

The estimates for mean log cotinine levels, their standard errors, and their relationship to the parameters are summarized in the table above. With cell mean coding, the parameter interpretations are clear; they each simply represent the mean log cotinine level for one task-wet combination.

part iii: The Full Model in Every Cell:

For these questions, use the log of salivary cotinine as the response and task and lnnsmoke as predictors.

```
tobacco3 <- tobacco %>% mutate(LOGCOT = log(COTININE),
                               TASK1 = case_when(TASK == 1 ~ 1, TASK != 1 ~ 0),
                               TASK2 = case_when(TASK == 2 ~ 1, TASK != 2 ~ 0),
                               TASK3 = case_when(TASK == 3 ~ 1, TASK != 3 ~ 0),
                               TASK4 = case_when(TASK == 4 ~ 1, TASK != 4 ~ 0))
```

(a):

Fit the full model in every cell. Provide and interpret estimates of all parameters for this model.

Consider the following model using the reference cell coding scheme:

$$y = \beta_1 + \beta_2 I_{T2} + \beta_3 I_{T3} + \beta_4 I_{T4} + \beta_5 X + \beta_6 I_{T2} X + \beta_7 I_{T3} X + \beta_8 I_{T4} X$$

where y is the log cotinine, X is the lnnsmoke varaible, and I_{Ti} is the indicator function associated with the ith task level. Note that TASK1 is the reference.

```
fit = lm(LOGCOT ~ factor(TASK) + factor(TASK)*LNNSMOKE, data = tobacco3)
summary(fit)$coefficients
```

```
##                          Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)             4.3344142 0.09319895  46.507113 2.802358e-214
## factor(TASK)2          -1.2210290 0.19086147  -6.397462  2.924652e-10
## factor(TASK)3          -2.3221216 0.13561455 -17.122953  5.432182e-55
## factor(TASK)4          -3.4207011 0.13356892 -25.610007 4.927483e-102
## LNNSMOKE                0.2945994 0.08869134   3.321625  9.424000e-04
## factor(TASK)2:LNNSMOKE  0.4274696 0.16993370   2.515508  1.211344e-02
## factor(TASK)3:LNNSMOKE  0.9359153 0.12592833   7.432126  3.191248e-13
## factor(TASK)4:LNNSMOKE  1.4843094 0.13531844  10.969010  6.613806e-26
```

The estimates for each parameter can be seen by the summary above. The intercept, β_1, is the mean log continine level for priming when lnnsmoke is 0. The factor(TASK)2 estimate, for β_2, is the difference between the mean log continine level for barning and for priming when lnnsmoke is 0. Similarly, The factor(TASK)2 estimate, for β_3, is the difference between the mean log continine level for topping and for priming when lnnsmoke is 0 and the factor(TASK)4 estimate, for β_4, is the difference between the mean log continine level for work not involving tobacco and for priming when lnnsmoke is 0. The LNNSMOKE estimate, for β_5, is the

mean increase log continine level for a one unit increase in lnnsmoke (the natural log of 1 + number of cigarettes smoked a day) in those whose task is priming. Similiarly, factor(TASK)i:LNNSMOKE variables, estimating $\beta_6, \beta_7, beta_8$, are the mean increase log continine level for a one unit increase in lnnsmoke for those whose task is barning, topping, and no tobacco involvement, respectively.

(b):

Report an appropriate test of whether task is related to cotinine levels. If this test is significant, report step-down tests to determine exactly where differences lie. For all tests reported, be sure to state H_0 clearly and give explicit justification for which tests were used and why.

In order to test whether task is realted to cotinine levels, we want to test the following hypotheses:

$$H_0 = 0 = \beta_2 = \beta_3 = \beta_4 \quad \& \quad 0 = \beta_6 = \beta_7 = \beta_8$$

which is equivalent to:

$$H_0 = \beta_2 = 0, \beta_3 = 0, \beta_4 = 0, \beta_6 = 0, \beta_7 = 0, \beta_8 = 0,$$

These will test whether the intercepts for each task and the slopes for each task are equivalent to each other. In order to test these hypotheses, we can use the overall F test, where:

$$F = (SSH/a)/\hat{\sigma}^2 \sim F_{G-2,n-G}$$

where $SSH = (\hat{\theta} - \theta_0)' M^{-1} (\hat{\theta} - \theta_0)$, $G = 8, n = 694$, and $^2 = $. It should also be noted that $M = C(X'X)^{-1}C'$.

```
X = tobacco3 %>% mutate(INT = 1, LNNTASK2 = LNNSMOKE*TASK2, LNNTASK3 = LNNSMOKE*TASK3, LNNTASK4 = LNNSMOKE*TASK4) %>% select(INT, TASK2, TASK
3, TASK4, LNNSMOKE, LNNTASK2, LNNTASK3, LNNTASK4) %>% as.matrix()
C = matrix(c(0, 0, 0, 0, 0, 0,
             1, 0, 0, 0, 0, 0,
             0, 1, 0, 0, 0, 0,
             0, 0, 1, 0, 0, 0,
             0, 0, 0, 0, 0, 0,
             0, 0, 0, 1, 0, 0,
             0, 0, 0, 0, 1, 0,
             0, 0, 0, 0, 0, 1), nrow = 6)

linearHypothesis(fit, C)
```

```
## Linear hypothesis test
##
## Hypothesis:
## factor(TASK)2 = 0
## factor(TASK)3 = 0
## factor(TASK)4 = 0
## factor(TASK)2:LNNSMOKE = 0
## factor(TASK)3:LNNSMOKE = 0
## factor(TASK)4:LNNSMOKE = 0
##
## Model 1: restricted model
## Model 2: LOGCOT ~ factor(TASK) + factor(TASK) * LNNSMOKE
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    692 1806.03
## 2    686  883.86  6    922.17 119.29 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the code summary above, the F statistic is 119.29 and the p-value is < 2.2e-16, so we can reject the null hypothesis. This means there is evidence to reject the fact the the slopes for all four task levels are equal and the intercepts for all four task levels are equal.

Next, we perform a step down test to test whether the intercepts are equal. The null hypothesis for this test will be the first set of hypotheses previously listed above.

```
C = matrix(c(0, 0, 0,
             1, 0, 0,
             0, 1, 0,
             0, 0, 1,
             0, 0, 0,
             0, 0, 0,
             0, 0, 0,
             0, 0, 0), nrow = 3)
linearHypothesis(fit, C)
```

```
## Linear hypothesis test
##
## Hypothesis:
## factor(TASK)2 = 0
## factor(TASK)3 = 0
## factor(TASK)4 = 0
##
## Model 1: restricted model
## Model 2: LOGCOT ~ factor(TASK) + factor(TASK) * LNNSMOKE
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    689 1785.88
## 2    686  883.86  3    902.01 233.36 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the code summary above, the F statistic is 233.36 and the p-value is < 2.2e-16, so we can reject the null hypothesis. This means there is evidence to reject the fact the the intercepts for all four task levels are equal.

We will now follow the same process with the slopes for each task level. The null hypothesis for this test will be the second set of hypotheses previously listed above at the first test for this problem part.

```
C = matrix(c(0, 0, 0,
             0, 0, 0,
             0, 0, 0,
             0, 0, 0,
             0, 0, 0,
             1, 0, 0,
             0, 1, 0,
             0, 0, 1), nrow = 3)

linearHypothesis(fit, C)
```

```
## Linear hypothesis test
##
## Hypothesis:
## factor(TASK)2:LNNSMOKE = 0
## factor(TASK)3:LNNSMOKE = 0
## factor(TASK)4:LNNSMOKE = 0
##
## Model 1: restricted model
## Model 2: LOGCOT ~ factor(TASK) + factor(TASK) * LNNSMOKE
##
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    689 1053.51
## 2    686  883.86  3    169.64 43.889 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the code summary above, the F statistic is 43.889 and the p-value is < 2.2e-16, so we can reject the null hypothesis. This means there is evidence to reject the fact the the slopes for all four task levels are equal.

Next, we would want to step down even further to test pairwise comparisons of the intercepts and of the slopes. For the intercepts, to test whether the intercepts for TASK1 and TASK2 are equivalent, we would want to test $H_0 : \beta_1 = \beta_1 + \beta_2 \rightarrow \beta_2 = 0$. To test the equivalence of TASK1 and TASK3, as well as TASK1 and TASK4, we can follow the same process. For TASK2 and TASK3, we would want to test $H_0 : \beta_2 = \beta_3 \rightarrow \beta_2 - \beta_3 = 0$. Similar hypotheses would be tested for TASK2 and TASK4, as well as TASK3 and TASK4. We will perform a normal F test, but will look to reject the null as p-values smaller than $\alpha = 0.05/6 = .008$ using the Bonferroni correction, since we are running 6 tests. All F-statistics have degrees of freedom (1, 686), since we are performing one test with a size $n - G$ fitted model.

```
### 1 and 2
C = matrix(c(0, 1, 0, 0, 0, 0, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f12 <- test$F[2]
p12 <- test$`Pr(>F)`[2]

### 1 and 3
C = matrix(c(0, 0, 1, 0, 0, 0, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f13 <- test$F[2]
p13 <- test$`Pr(>F)`[2]

### 1 and 4
C = matrix(c(0, 0, 0, 1, 0, 0, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f14 <- test$F[2]
p14 <- test$`Pr(>F)`[2]

### 2 and 3
C = matrix(c(0, 1, -1, 0, 0, 0, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f23 <- test$F[2]
p23 <- test$`Pr(>F)`[2]

### 2 and 4
C = matrix(c(0, 1, 0, -1, 0, 0, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f24 <- test$F[2]
p24 <- test$`Pr(>F)`[2]

### 3 and 4
C = matrix(c(0, 0, 1, -1, 0, 0, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f34 <- test$F[2]
p34 <- test$`Pr(>F)`[2]

df <- data.frame(Test = c("TASK1:TASK2", "TASK1:TASK3", "TASK1:TASK4", "TASK2:TASK3", "TASK2:TASK4", "TASK3:TASK4"), Fvalue = c(f12, f13, f14
, f23, f24, f34), Pvalue = c(p12, p13, p14, p23, p24, p34))
df %>% knitr::kable(align = c("c", "c"), digits = c(10,  10, 20))
```

Test	Fvalue	Pvalue
TASK1:TASK2	40.92752	2.924652e-10
TASK1:TASK3	293.19551	0.000000e+00
TASK1:TASK4	655.87245	0.000000e+00
TASK2:TASK3	32.37628	1.882119e-08
TASK2:TASK4	131.13804	0.000000e+00
TASK3:TASK4	63.99177	5.325440e-15

From the summary table above, we can reject the null hypothesis in every case. That is, there is evidence to support the fact that all combinations of TASK intercepts are significnantly different than one another. We will perform an analogous analysis for the pairwise testing of the equivalence of slopes of the TASK levels.

```
### 1 and 2
C = matrix(c(0, 0, 0, 0, 0, 1, 0, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f12 <- test$F[2]
p12 <- test$`Pr(>F)`[2]

### 1 and 3
C = matrix(c(0, 0, 0, 0, 0, 0, 1, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f13 <- test$F[2]
p13 <- test$`Pr(>F)`[2]

### 1 and 4
C = matrix(c(0, 0, 0, 0, 0, 0, 0, 1), nrow = 1)
test<-linearHypothesis(fit, C)
f14 <- test$F[2]
p14 <- test$`Pr(>F)`[2]

### 2 and 3
C = matrix(c(0, 0, 0, 0, 0, 1, -1, 0), nrow = 1)
test<-linearHypothesis(fit, C)
f23 <- test$F[2]
p23 <- test$`Pr(>F)`[2]

### 2 and 4
C = matrix(c(0, 0, 0, 0, 0, 1, 0, -1), nrow = 1)
test<-linearHypothesis(fit, C)
f24 <- test$F[2]
p24 <- test$`Pr(>F)`[2]

### 3 and 4
C = matrix(c(0, 0, 0, 0, 0, 0, 1, -1), nrow = 1)
test<-linearHypothesis(fit, C)
f34 <- test$F[2]
p34 <- test$`Pr(>F)`[2]

df <- data.frame(Test = c("TASK1:TASK2", "TASK1:TASK3", "TASK1:TASK4", "TASK2:TASK3", "TASK2:TASK4", "TASK3:TASK4"), Fvalue = c(f12, f13, f14
, f23, f24, f34), Pvalue = c(p12, p13, p14, p23, p24, p34))
df %>% knitr::kable(align = c("c", "c"), digits = c(10,  10, 20))
```

Test	Fvalue	Pvalue
TASK1:TASK2	6.327780	1.211344e-02
TASK1:TASK3	55.236500	3.191248e-13
TASK1:TASK4	120.319184	0.000000e+00
TASK2:TASK3	8.913427	2.931579e-03
TASK2:TASK4	35.506803	4.063003e-09
TASK3:TASK4	16.311803	5.979385e-05

From the summary table above, we can reject the null hypothesis in almost every case. That is, there is evidence to support the fact that all combinations of TASK slopes (except TASK1:TASK2) are significnantly different than one another. the p-value for TASK1:TASK2 is greater than the Bonferroni corrected $\alpha = 0.008$ so there is not evidence to reject the fact that the mean increase in log cotinine levels when lnnsmoke is increased by one is different between priming and barning tasks.

```
ods pdf file = "/home/sedwards/UNC/B663_H4.pdf";
ods graphics on;

LIBNAME home "/home/sedwards/UNC/663/";
%LET filepath = /home/sedwards/UNC/663/;

/**** Question #1 *****
172 young adult males received a battery of pulmonary function tests.

Fit a model with average forced vital capacity (FVC) (in ml) as the outcome
with height, weight, body mass index, age, average treadmill elivation,
      average treadmill speed, temperature, barometric pressure, and humidity
      as predictors.

Consider the model -
  FVC = B0 + B1(height) + B2(weight) + B3(BMI) + B4(area) + B5(age) +
  B6(avtrel) + B7(avtrsp)
          + B8(avtrel)(avtrsp) + B9(temp) + B10(barm) + B11(hum) + e

***/
TITLE "QUESTION 1";
/* READ IN DATA & SET UP VARIABLES */
DATA filen;
    INFILE "&filepath.\FILEN.DAT";
    INPUT subject year cohort date days timsess height weight age area temp
    barm
            hum avtrel avtrsp avfvc;
    LABEL
    subject="subject id"
    year="year of study"
    cohort="Ozone Dosage Level Group"
    date="Date of Study, Julian Date"
    days="# Days After 12-31-79"
    timsess="Time of Session"
    height="Height (cm)"
    weight="Weight (kg)"
    age="Age (years)"
    area="Body Surface Area (M**2)"
    temp="Air Temperature (deg C)"
    barm="Barometric Pressure (mmHg)"
    hum="Relative Humidity %"
    avtrel="Average Treadmill Elevation (deg)"
    avtrsp="Average Speed of Treadmill (mph)"
    avfvc="Average Forced Vital Capacity (mL)";
RUN;

DATA filen;
    SET filen;

    int=avtrel*avtrsp;
    bmi=10000*weight/(height*height);
    tim2=timsess*timsess;
RUN;


PROC FREQ DATA=filen;
    TABLES height weight BMI area avtrel avtrsp int temp barm hum age / LIST
    MISSING;
RUN;

PROC FREQ DATA=filen;
    WHERE weight=.;
    TABLES weight*bmi / LIST MISSING;
RUN;

/***
```

```
(a) Examine the tolerances and variance inflation factors (VIF) from this
model.
    Do you think any collinearity is present based on the tolerance and VIF?
    Why or why not?
***/
PROC REG DATA=filen;
    MODEL avfvc = height weight BMI area age avtrel avtrsp int temp barm hum
    / TOL VIF;
    TITLE "Q1-A";
QUIT;

/***
(b) Conduct an eigenanalysis of the scaled SSCP and correlation matrices,
presenting a table formatted
    like Table 8.6.2.

  i) Does there appear to be collinarity with the intercept? Why or why not?
     If so which variables are suspect?

 ii) Does there appear to be any other collinearity? Why or why not?
     If so which variables are suspect?
***/
DATA B;
    SET filen;
    inter = 1;
RUN;
ods output Eigenvalues=EIa;
PROC PRINCOMP DATA=B NOINT;     /* SCALED SSCP MATRIX */
    VAR inter height weight BMI area age avtrel avtrsp int temp barm hum;
    TITLE "Q1-B";
RUN;

DATA EIa;
    IF _N_ = 1 THEN DO; SET EIa(RENAME=(eigenvalue=e1)); WHERE number = 1; END;
    SET EIa;
    CondIndex = sqrt(e1/eigenvalue);
RUN;
PROC PRINT DATA=EIa;    RUN;

ods output Eigenvalues=EIb;
PROC PRINCOMP DATA=B;           /* CORRELATION MATRIX */
    VAR inter height weight BMI area age avtrel avtrsp int temp barm hum;
RUN;
DATA EIb;
    IF _N_ = 1 THEN DO; SET EIb(RENAME=(eigenvalue=e1)); WHERE number = 1; END;
    SET EIb;
    CondIndex = sqrt(e1/eigenvalue);
RUN;
PROC PRINT DATA=EIb; RUN;


/**** Question #2 *****
Find the Box-Cox transformation of the simulated data (boxcox.dat) and
  compare the residual plots of the raw and transformed data.
***/
TITLE "QUESTION 2";
/* READ IN DATA & SET UP VARIABLES */
DATA filen;
    INFILE "&filepath.BoxCox.dat";
    INPUT x y;
RUN;

/* BEST LAMBDA = 0.5 */
PROC TRANSREG DATA=filen SS2 DETAIL;
    TITLE "DEFAULTS";
    MODEL BOXCOX(y) = IDENTITY(x);
```

```
RUN;

PROC IML;
USE filen;
READ ALL VAR "y" INTO y;
lny=log(y);
n=nrow(y);
one=j(n,1,1);
avglog=lny`*one/n;
PRINT avglog;
geomean = exp(avglog);
PRINT geomean;
geomeany=geomean#one;
create gmean var {geomeany};
append from geomeany;
close gmean;
QUIT;
RUN;

DATA p1;
    MERGE gmean filen;

    y_5 = ((y**0.5)-1)/(0.5*(geomeany**(0.5-1)));
RUN;

ODS GRAPHICS ON;
PROC REG DATA=p1 PLOTS=ALL;
    MODEL y = x;
    TITLE "1-RAW";
QUIT;

PROC REG DATA=p1 PLOTS=ALL;
    MODEL y_5 = x;
    TITLE "1-TRANSFORMED";
QUIT;
ODS GRAPHICS OFF;


/**** Question #3 *****
Effect of dermal nicotine exposure in a population of Latino tobacco workers
  in North Carolina.
***/
TITLE "QUESTION 3";

/* READ IN DATA & SET UP VARIABLES */
DATA filen;
    INFILE "&filepath.\tobacco.dat";
    INPUT cotinine age bmi educ wet task lnnsmoke;

    ID = _N_;
    lnCot = LOG(cotinine);

    IF task=1 THEN t1=1; ELSE t1=0;
    IF task=2 THEN t2=1; ELSE t2=0;
    IF task=3 THEN t3=1; ELSE t3=0;
    IF task=4 THEN t4=1; ELSE t4=0;

    wet_0 = 1 - wet;
    wet_1 = wet;

    w0t1 = wet_0 AND t1;
    w0t2 = wet_0 AND t2;
    w0t3 = wet_0 AND t3;
    w0t4 = wet_0 AND t4;
    w1t1 = wet_1 AND t1;
    w1t2 = wet_1 AND t2;
```

```
191       w1t3 = wet_1 AND t3;
192       w1t4 = wet_1 AND t4;
193
194       lnsmkt1 = t1 * lnnsmoke;
195       lnsmkt2 = t2 * lnnsmoke;
196       lnsmkt3 = t3 * lnnsmoke;
197       lnsmkt4 = t4 * lnnsmoke;
198
199       one = 1;
200   RUN;
201
202   PROC MEANS DATA=filen;
203       VAR cotinine age bmi educ wet task lnnsmoke;
204   RUN;
205
206   PROC FREQ DATA=filen;
207       TABLES age educ wet task / LIST MISSING;
208   RUN;
209
210   PROC SORT DATA=filen; BY cotinine; RUN;
211   PROC PRINT DATA=filen (FIRSTOBS=664);   RUN;
212
213
214   /***
215   3A - ONE-WAY ANOVA - LOG(cotinine) = task
216     Are all cell means equal? - H0/df/p-value/decision/interpretation
217     Examine all pairwise comparisons using the Scheffe correction.
218       Summerize findings in a table - estimated mean diff/df/F/p-value/Scheffe
          CI for mean diff
219       Explain your findings in language the investigator can understand.
220     Create a table of parameter estimates and standard errors - cell mean
        coding / reference cell coding
221       - give interpretations of parameters in both
222       - provide C and theta matrices for testing avg. cotinine (task=1) > avg.
          cotinine (task=234)
223   ***/
224   PROC GLM DATA=filen;
225       MODEL lnCot = t1 t2 t3 t4 / NOINT;
226       CONTRAST "Usual Overall Test" t1 1 t2 -1 t3  0 t4  0,
227                                     t1 1 t2 0  t3 -1 t4  0,
228                                     t1 1 t2 0  t3  0 t4 -1,
229                                     t1 0 t2 1  t3 -1 t4  0,
230                                     t1 0 t2 1  t3  0 t4 -1;
231
232       CONTRAST "t1 v t2" t1 1 t2 -1 t3  0 t4  0;
233       CONTRAST "t1 v t3" t1 1 t2  0 t3 -1 t4  0;
234       CONTRAST "t1 v t4" t1 1 t2  0 t3  0 t4 -1;
235       CONTRAST "t2 v t3" t1 0 t2  1 t3 -1 t4  0;
236       CONTRAST "t2 v t4" t1 0 t2  1 t3  0 t4 -1;
237       CONTRAST "t3 v t4" t1 0 t2  0 t3  1 t4 -1;
238
239       CONTRAST "T1 > AVG(T234)" t1 1 t2 -0.333333 t3 -0.333333 t4 -0.333333;
240
241       TITLE "CELL MEANS";
242   QUIT;
243
244   PROC GLM DATA=filen;
245       CLASS task(REF=LAST);
246       MODEL lnCot = task / SOLUTION;
247       TITLE "REF CELL - 1";
248   QUIT;
249
250   PROC GLM DATA=filen;
251       MODEL lnCot = one t1 t2 t3 / NOINT SOLUTION;
252       CONTRAST "Usual Overall Test" one 0 t1 1 t2 0 t3 0,
253                                     one 0 t1 0 t2 1 t3 0,
```

```
                                    one 0 t1 0 t2 0 t3 1;
    CONTRAST "T1 > AVG(T234)" one 0 t1 1 t2 -0.333333 t3 -0.333333;

    TITLE "REF CELL - 2";
QUIT;

PROC GLM DATA=filen;
    CLASS task;
    MODEL lnCot = task / NOINT;
    LSMEANS task / PDIFF ADJUST=SCHEFFE;
    MEANS task / SCHEFFE;

    TITLE "SCHEFFE";
QUIT;


/***
B - TWO-WAY ANOVA - LOG(cotinine) = task + wet

lnCot = mu + alpha i + beta j + gamma ij

i=2     j=4     ij=8

1 + 2 + 4 + 8 = 15

let wet=1 and task=4 be the reference
i=1
j=3     1 + 1 + 3 + 3 = 8

A(mu) + B(wet_0) + C(t1) + D(t2) + E(t3) + BC(w0t1) + BD (w0t2) + BE (w0t3)

now 4 parameters describe the dose effect

  Fit model with full interaction.
    - Interpret all parameter estimates.
    - Clearly state coding scheme used.
    - Discuss HILE-G assumptions.
  Create a table of estimated mean log(cotinine) levels, standard errors, how
  each estaimated mean is obtained from the model parameters
***/
PROC FREQ DATA=filen;
    TABLES task*wet / MISSING NOROW NOCOL NOPERCENT;
    TITLE "Balanced Cells?";
RUN;

PROC GLM DATA=filen;
    MODEL lnCot = w0t1 w1t1 w0t2 w1t2 w0t3 w1t3 w0t4 w1t4 / NOINT SOLUTION;

    ESTIMATE "GRAND MEAN" w0t1 1 w1t1 1 w0t2 1 w1t2 1 w0t3 1 w1t3 1 w0t4 1
    w1t4 1 / DIVISOR=8;
    ESTIMATE "MARG MEAN: WET 0" w0t1 1 w1t1 0 w0t2 1 w1t2 0 w0t3 1 w1t3 0
    w0t4 1 w1t4 0 / DIVISOR=4;
    ESTIMATE "MARG MEAN: WET 1" w0t1 0 w1t1 1 w0t2 0 w1t2 1 w0t3 0 w1t3 1
    w0t4 0 w1t4 1 / DIVISOR=4;
    ESTIMATE "MARG MEAN: T1" w0t1 1 w1t1 1 w0t2 0 w1t2 0 w0t3 0 w1t3 0 w0t4 0
    w1t4 0 / DIVISOR=2;
    ESTIMATE "MARG MEAN: T2" w0t1 0 w1t1 0 w0t2 1 w1t2 1 w0t3 0 w1t3 0 w0t4 0
    w1t4 0 / DIVISOR=2;
    ESTIMATE "MARG MEAN: T3" w0t1 0 w1t1 0 w0t2 0 w1t2 0 w0t3 1 w1t3 1 w0t4 0
    w1t4 0 / DIVISOR=2;
    ESTIMATE "MARG MEAN: T4" w0t1 0 w1t1 0 w0t2 0 w1t2 0 w0t3 0 w1t3 0 w0t4 1
    w1t4 1 / DIVISOR=2;

    CONTRAST "MAIN EFFECT WET" w0t1 1 w1t1 -1 w0t2 1 w1t2 -1 w0t3 1 w1t3 -1
    w0t4 1 w1t4 -1;

```

```
311       CONTRAST "MAIN EFFECT TASK" w0t1 1 w1t1 1 w0t2 -1 w1t2 -1 w0t3  0 w1t3  0
          w0t4  0 w1t4  0,
312                                   w0t1 1 w1t1 1 w0t2  0 w1t2  0 w0t3 -1 w1t3 -1
                                      w0t4  0 w1t4  0,
313                                   w0t1 1 w1t1 1 w0t2  0 w1t2  0 w0t3  0 w1t3  0
                                      w0t4 -1 w1t4 -1;
314
315       CONTRAST "INTERACTION WET V TASK" w0t1 1 w1t1 -1 w0t2 -1 w1t2  1 w0t3  0
          w1t3  0 w0t4  0 w1t4  0,
316                                         w0t1 1 w1t1 -1 w0t2  0 w1t2  0 w0t3 -1
                                            w1t3  1 w0t4  0 w1t4  0,
317                                         w0t1 1 w1t1 -1 w0t2  0 w1t2  0 w0t3  0
                                            w1t3  0 w0t4 -1 w1t4  1;
318
319       TITLE "CELL MEAN";
320   QUIT;
321
322   PROC GLM DATA=filen;
323       MODEL lnCot = one wet_1 t2 t3 t4 w1t2 w1t3 w1t4 / NOINT SOLUTION;
324
325       ESTIMATE "GRAND MEAN" one 8 wet_1 4 t2 2 t3 2 t4 2 w1t2 1 w1t3 1 w1t4 1 /
          DIVISOR=8;
326       ESTIMATE "MARG MEAN: WET 0" one 4 wet_1 0 t2 1 t3 1 t4 1 w1t2 0 w1t3 0
          w1t4 0 / DIVISOR=4;
327       ESTIMATE "MARG MEAN: WET 1" one 4 wet_1 4 t2 1 t3 1 t4 1 w1t2 1 w1t3 1
          w1t4 1 / DIVISOR=4;
328       ESTIMATE "MARG MEAN: T1" one 2 wet_1 1 t2 0 t3 0 t4 0 w1t2 0 w1t3 0 w1t4
          0 / DIVISOR=2;
329       ESTIMATE "MARG MEAN: T2" one 2 wet_1 1 t2 2 t3 0 t4 0 w1t2 1 w1t3 0 w1t4
          0 / DIVISOR=2;
330       ESTIMATE "MARG MEAN: T3" one 2 wet_1 1 t2 0 t3 2 t4 0 w1t2 0 w1t3 1 w1t4
          0 / DIVISOR=2;
331       ESTIMATE "MARG MEAN: T4" one 2 wet_1 1 t2 0 t3 0 t4 2 w1t2 0 w1t3 0 w1t4
          1 / DIVISOR=2;
332
333       CONTRAST "MAIN EFFECT WET" one 0 wet_1 4 t2 0 t3 0 t4 0 w1t2 1 w1t3 1
          w1t4 1;
334
335       CONTRAST "MAIN EFFECT TASK" one 0 wet_1 0 t2 2 t3 0 t4 0 w1t2 1 w1t3 0
          w1t4 0,
336                                   one 0 wet_1 0 t2 0 t3 2 t4 0 w1t2 0 w1t3 1
                                      w1t4 0,
337                                   one 0 wet_1 0 t2 0 t3 0 t4 2 w1t2 0 w1t3 0
                                      w1t4 1;
338
339       CONTRAST "INTERACTION WET V TASK" one 0 wet_1 0 t2 0 t3 0 t4 0 w1t2 1
          w1t3 0 w1t4 0,
340                                         one 0 wet_1 0 t2 0 t3 0 t4 0 w1t2 0
                                            w1t3 1 w1t4 0,
341                                         one 0 wet_1 0 t2 0 t3 0 t4 0 w1t2 0
                                            w1t3 0 w1t4 1;
342
343       TITLE "REF CELL";
344   QUIT;
345
346   ODS GRAPHICS ON;
347   PROC REG DATA=filen PLOTS=ALL;
348       MODEL lnCot = one wet_1 t2 t3 t4 w1t2 w1t3 w1t4 / NOINT;
349       TITLE "3-B";
350       OUTPUT OUT=B PREDICTED=y_hat RSTUDENT=r_i;
351   QUIT;
352
353   PROC UNIVARIATE DATA=B PLOT NORMAL;
354       CLASS wet task;
355       VAR r_i;
356       QQPLOT r_i / NORMAL;
```

```
357       HISTOGRAM r_i / NORMAL;
358       TITLE "3-B";
359   RUN;
360
361   PROC MEANS DATA=B STD;
362       CLASS y_hat;
363       VAR r_i;
364       OUTPUT OUT=B_2 STD(r_i)=sd;
365   RUN;
366
367   PROC SGPLOT DATA=B_2;
368       SCATTER X=y_hat Y=sd;
369       label sd="Within Group Sample SDs";
370       TITLE "Within Group Residuals";
371   RUN;
372
373   ODS GRAPHICS OFF;
374
375
376   /***
377   C - FULL MODEL IN EVERY CELL - LOG(cotinine) = task + lnnsmoke - categorical
      & continuous
378
379     Fit full model in every cell.
380       - Interpret all parameter estimates.
381     Report test for whether task is related to cotinine levels.
382       - If sig, report step-down tests to determine exactly where differences
          lie.
383       - State H0 & Give justification for which test is used and why.
384   ***/
385   PROC MEANS DATA=filen;
386       VAR lnnsmoke;           /* mean = 0.5960 */
387   RUN;
388
389   PROC GLM DATA=filen;
390       MODEL lnCot = t1 t2 t3 t4 lnsmkt1 lnsmkt2 lnsmkt3 lnsmkt4 / NOINT SOLUTION;
391
392       ESTIMATE "ADJ CELL MEAN: T1" t1 1 t2 0 t3 0 t4 0 lnsmkt1 0.5960 lnsmkt2 0
          lnsmkt3 0 lnsmkt4 0;
393       ESTIMATE "ADJ CELL MEAN: T2" t1 0 t2 1 t3 0 t4 0 lnsmkt1 0 lnsmkt2 0.5960
          lnsmkt3 0 lnsmkt4 0;
394       ESTIMATE "ADJ CELL MEAN: T3" t1 0 t2 0 t3 1 t4 0 lnsmkt1 0 lnsmkt2 0
          lnsmkt3 0.5960 lnsmkt4 0;
395       ESTIMATE "ADJ CELL MEAN: T4" t1 0 t2 0 t3 0 t4 1 lnsmkt1 0 lnsmkt2 0
          lnsmkt3 0 lnsmkt4 0.5960;
396
397       ESTIMATE "MEAN OF ADJ CELL MEANS" t1 1 t2 1 t3 1 t4 1 lnsmkt1 0.5960
          lnsmkt2 0.5960 lnsmkt3 0.5960 lnsmkt4 0.5960 / DIVISOR=4;
398       ESTIMATE "MEAN INTERCEPT" t1 1 t2 1 t3 1 t4 1 lnsmkt1 0 lnsmkt2 0 lnsmkt3
          0 lnsmkt4 0 / DIVISOR=4;
399       ESTIMATE "MEAN SLOPE" t1 0 t2 0 t3 0 t4 0 lnsmkt1 1 lnsmkt2 1 lnsmkt3 1
          lnsmkt4 1 / DIVISOR=4;
400
401       CONTRAST "TEST OF COINCIDENCE" t1 1 t2 -1 t3  0 t4  0 lnsmkt1  0 lnsmkt2
          0 lnsmkt3  0 lnsmkt4  0,
402                                      t1 1 t2  0 t3 -1 t4  0 lnsmkt1  0 lnsmkt2
                                         0 lnsmkt3  0 lnsmkt4  0,
403                                      t1 1 t2  0 t3  0 t4 -1 lnsmkt1  0 lnsmkt2
                                         0 lnsmkt3  0 lnsmkt4  0,
404                                      t1 0 t2  0 t3  0 t4  0 lnsmkt1  1 lnsmkt2
                                         -1 lnsmkt3  0 lnsmkt4  0,
405                                      t1 0 t2  0 t3  0 t4  0 lnsmkt1  1 lnsmkt2
                                         0 lnsmkt3 -1 lnsmkt4  0,
406                                      t1 0 t2  0 t3  0 t4  0 lnsmkt1  1 lnsmkt2
                                         0 lnsmkt3  0 lnsmkt4 -1;
407
```

```
408       CONTRAST "STEPDOWN: EQUAL INTERCEPTS"
409                                    t1 1 t2 -1 t3  0 t4  0 lnsmkt1  0 lnsmkt2
                                       0 lnsmkt3  0 lnsmkt4  0,
410                                    t1 1 t2  0 t3 -1 t4  0 lnsmkt1  0 lnsmkt2
                                       0 lnsmkt3  0 lnsmkt4  0,
411                                    t1 1 t2  0 t3  0 t4 -1 lnsmkt1  0 lnsmkt2
                                       0 lnsmkt3  0 lnsmkt4  0;
412       CONTRAST "STEPDOWN: EQUAL SLOPES"
413                                    t1 0 t2  0 t3  0 t4  0 lnsmkt1  1 lnsmkt2
                                       -1 lnsmkt3  0 lnsmkt4  0,
414                                    t1 0 t2  0 t3  0 t4  0 lnsmkt1  1 lnsmkt2
                                       0 lnsmkt3 -1 lnsmkt4  0,
415                                    t1 0 t2  0 t3  0 t4  0 lnsmkt1  1 lnsmkt2
                                       0 lnsmkt3  0 lnsmkt4 -1;
416
417       CONTRAST "PAIRWISE INTERCEPTS T1 V T2" t1 1 t2 -1 t3  0 t4  0 lnsmkt1  0
          lnsmkt2  0 lnsmkt3  0 lnsmkt4  0;
418       CONTRAST "PAIRWISE INTERCEPTS T1 V T3" t1 1 t2  0 t3 -1 t4  0 lnsmkt1  0
          lnsmkt2  0 lnsmkt3  0 lnsmkt4  0;
419       CONTRAST "PAIRWISE INTERCEPTS T1 V T4" t1 1 t2  0 t3  0 t4 -1 lnsmkt1  0
          lnsmkt2  0 lnsmkt3  0 lnsmkt4  0;
420       CONTRAST "PAIRWISE INTERCEPTS T2 V T3" t1 0 t2  1 t3 -1 t4  0 lnsmkt1  0
          lnsmkt2  0 lnsmkt3  0 lnsmkt4  0;
421       CONTRAST "PAIRWISE INTERCEPTS T2 V T4" t1 0 t2  1 t3  0 t4 -1 lnsmkt1  0
          lnsmkt2  0 lnsmkt3  0 lnsmkt4  0;
422       CONTRAST "PAIRWISE INTERCEPTS T3 V T4" t1 0 t2  0 t3  1 t4 -1 lnsmkt1  0
          lnsmkt2  0 lnsmkt3  0 lnsmkt4  0;
423
424       CONTRAST "PAIRWISE SLOPES T1 V T2" t1 0 t2  0 t3  0 t4  0 lnsmkt1  1
          lnsmkt2 -1 lnsmkt3  0 lnsmkt4  0;
425       CONTRAST "PAIRWISE SLOPES T1 V T3" t1 0 t2  0 t3  0 t4  0 lnsmkt1  1
          lnsmkt2  0 lnsmkt3 -1 lnsmkt4  0;
426       CONTRAST "PAIRWISE SLOPES T1 V T4" t1 0 t2  0 t3  0 t4  0 lnsmkt1  1
          lnsmkt2  0 lnsmkt3  0 lnsmkt4 -1;
427       CONTRAST "PAIRWISE SLOPES T2 V T3" t1 0 t2  0 t3  0 t4  0 lnsmkt1  0
          lnsmkt2  1 lnsmkt3 -1 lnsmkt4  0;
428       CONTRAST "PAIRWISE SLOPES T2 V T4" t1 0 t2  0 t3  0 t4  0 lnsmkt1  0
          lnsmkt2  1 lnsmkt3  0 lnsmkt4 -1;
429       CONTRAST "PAIRWISE SLOPES T3 V T4" t1 0 t2  0 t3  0 t4  0 lnsmkt1  0
          lnsmkt2  0 lnsmkt3  1 lnsmkt4 -1;
430
431       CONTRAST "EQUAL ADJ CELL MEANS" t1 1 t2 -1 t3  0 t4  0 lnsmkt1  0.5960
          lnsmkt2  -0.5960 lnsmkt3  0 lnsmkt4  0,
432                                     t1 1 t2  0 t3 -1 t4  0 lnsmkt1  0.5960
                                        lnsmkt2  0 lnsmkt3  -0.5960 lnsmkt4  0,
433                                     t1 1 t2  0 t3  0 t4 -1 lnsmkt1  0
                                        lnsmkt2  0 lnsmkt3  0 lnsmkt4  -0.5960;
434       CONTRAST "PAIRWISE ADJ T1 V T2" t1 1 t2 -1 t3  0 t4  0 lnsmkt1  0.5960
          lnsmkt2  -0.5960 lnsmkt3 0 lnsmkt4  0;
435       CONTRAST "PAIRWISE ADJ T1 V T3" t1 1 t2  0 t3 -1 t4  0 lnsmkt1  0.5960
          lnsmkt2  0 lnsmkt3 -0.5960 lnsmkt4  0;
436       CONTRAST "PAIRWISE ADJ T1 V T4" t1 1 t2  0 t3  0 t4 -1 lnsmkt1  0.5960
          lnsmkt2  0 lnsmkt3  0 lnsmkt4 -0.5960;
437       CONTRAST "PAIRWISE ADJ T2 V T3" t1 0 t2  1 t3 -1 t4  0 lnsmkt1  0
          lnsmkt2  0.5960 lnsmkt3 -0.5960 lnsmkt4  0;
438       CONTRAST "PAIRWISE ADJ T2 V T4" t1 0 t2  1 t3  0 t4 -1 lnsmkt1  0
          lnsmkt2  0.5960 lnsmkt3  0 lnsmkt4 -0.5960;
439       CONTRAST "PAIRWISE ADJ T3 V T4" t1 0 t2  0 t3  1 t4 -1 lnsmkt1  0
          lnsmkt2  0 lnsmkt3  0.5960 lnsmkt4 -0.5960;
440
441       TITLE "3-C CELL MEANS";
442   QUIT;
443
444   PROC GLM DATA=filen;
445       CLASS task;
446       MODEL lnCot = task lnnsmoke / NOINT;
```

```
447       MEANS task / SCHEFFE;
448       LSMEANS task / PDIFF ADJUST=SCHEFFE;
449       TITLE "SCHEFFE";
450   QUIT;
451
452
```

1. The following questions are on the data and model described in Q3 of HW3:
   - Examine the tolerances and variance inflation factors from this model. Do you think any collinearity is present based on the tolerance and VIF? Why or why not?

Table 1.1: Tolerances & Variance Inflation Factors

Variable	Label	DF	Tolerance	Variance Inflation
Intercept	Intercept	1	.	0
height	Height (cm)	1	0.00218	458.04764
weight	Weight (kg)	1	0.00142	703.44629
bmi		1	0.00564	177.45037
area	Body Surface Area (M**2)	1	0.00073266	1364.89752
age	Age (years)	1	0.92307	1.08334
avtrel	Average Treadmill Elevation (deg)	1	0.00172	580.38969
avtrsp	Average Speed of Treadmill (mph)	1	0.01276	78.39302
int		1	0.00126	795.44895
temp	Air Temperature (deg C)	1	0.03447	29.01379
barm	Barometric Pressure (mmHg)	1	0.94420	1.05910
hum	Relative Humidity %	1	0.03429	29.16692

- There is a good amount of evidence to suggest collinearity in:
  - BMI and body surface area can be predicted/derived from height and weight.
  - Average treadmill elevation, average treadmill speed, and their interaction are collinear.
  - Temperature and relative humidity can be somewhat "predicted" once one of them is known.
- Age and barometric pressure do not appear to be collinear. This makes sense since age cannot guarantee height and weight among adults. Barometric pressure would probably tend to stay constant or within a narrow range of values while humidity can vary much more.

- Conduct an eigenanalysis of the scaled SSCP and correlation matrices, presenting a table formatted like Table 8.6.2 in Muller and Fetterman.
  (a) Does there appear to be any collinearity between the intercept and the covariates? Why or why not? If so, list the variables?

Table 1.2: Eigenvalues & Condition Index – Scaled SSCP Matrix

Number	Eigenvalue	CondIndex
1	11.9049480	1.00
2	0.0360383	18.18
3	0.0293709	20.13
4	0.0161788	27.13
5	0.0066991	42.16
6	0.0049049	49.27
7	0.0015550	87.50
8	0.0002515	217.55
9	0.0000375	563.68
10	0.0000097	1109.52
11	0.0000049	1563.90
12	0.0000015	2772.26

Table 1.3: Eigenvectors – Scaled SSCP Matrix

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
inter	0.289612	0.003847	-.041736	0.011264	-.428489	0.169349
height	0.289592	-.012406	-.076762	-.085857	-.320422	-.333604
weight	0.288434	-.239953	-.360268	-.362654	0.422414	-.270732
bmi	0.288728	-.206145	-.292000	-.169993	0.215099	0.744305
area	0.289363	-.111646	-.201362	-.220653	0.010005	-.386236
age	0.287541	-.065062	-.310877	0.876080	0.184729	-.112124
avtrel	0.288114	0.542971	0.114554	-.008826	0.189245	0.156326
avtrsp	0.289555	0.081696	-.006435	-.063918	-.344231	-.063169
int	0.287589	0.621228	0.147453	-.079888	0.268208	-.082800
temp	0.288619	-.262600	0.436206	0.056399	0.018763	0.043500
barm	0.289608	0.005253	-.047742	0.006020	-.425070	0.178686
hum	0.287332	-.355890	0.642383	0.046041	0.217990	-.043461

	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12
inter	-.160576	0.078500	-.338231	-.161589	0.611844	-.393890
height	-.210883	0.038668	-.291386	0.338828	0.023636	0.665223
weight	-.018722	-.039185	0.238215	-.394929	0.334969	0.139220
bmi	0.123098	0.027757	-.194962	0.269726	-.143594	0.111429
area	-.132498	0.015069	-.118058	0.290762	-.440686	-.595672
age	0.045774	0.011125	0.001522	-.001560	-.002277	-.001034
avtrel	-.507446	0.047211	-.121532	-.414212	-.312294	0.071559
avtrsp	0.689454	-.012728	-.115575	-.438188	-.311166	0.071916
int	0.359344	-.055730	0.120806	0.416970	0.313806	-.072700
temp	-.078823	-.802916	0.031595	0.002606	-.009520	-.007541
barm	-.155432	0.124637	0.804979	0.091474	-.071280	0.006178
hum	0.047328	0.569193	-.016930	0.000288	0.007634	0.005636

Several of the scaled SSCP condition indices are greater than 30 and multiple eigenvalues are close to zero; thus, indicating collinarity with the intercept.

By looking at the elements of the 12th Prinicipal Component, the elements corresponding to the intercept, height, weight, bmi, and area have values father from zero relative to the other variables.

This would suggest that height, weight, bmi, and area span the intercept.

(b) Does there appear to be any collinearity among the covariates? Why or why not? If so, list the variables?

Table 1.4: Eigenvalues & Condition Index - Correlation

Number	Eigenvalue	CondIndex
1	3.00984215	1.0000
2	2.44782689	1.1089
3	2.02476481	1.2192
4	1.11320127	1.6443
5	1.01325013	1.7235
6	0.80927943	1.9285
7	0.56109511	2.3161
8	0.01770758	13.0374
9	0.00187374	40.0791
10	0.00070517	65.3317
11	0.00045372	81.4475

Table 1.5: Eigenvectors – Correlation

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6
height	0.392316	0.296786	0.032498	-.431698	0.052791	0.294700
weight	0.558767	0.051052	-.092125	0.096942	0.000376	-.170215
bmi	0.367229	-.183577	-.143312	0.493420	-.041432	-.472648
area	0.549273	0.157720	-.048934	-.111776	0.019229	-.000896
age	0.092890	-.065940	-.101323	0.368850	0.755491	0.497666
avtrel	-.199140	0.514774	-.034223	0.298870	0.116311	-.185008
avtrsp	0.074811	0.479606	0.086722	-.079394	-.115421	0.054060
int	-.148534	0.589598	-.004523	0.228684	0.068080	-.138848
temp	0.091766	-.044660	0.676385	0.168590	-.008712	0.038468
barm	0.065437	0.035488	-.176120	0.457649	-.626029	0.594893
hum	0.093072	-.022211	0.678255	0.162613	-.029775	0.036111

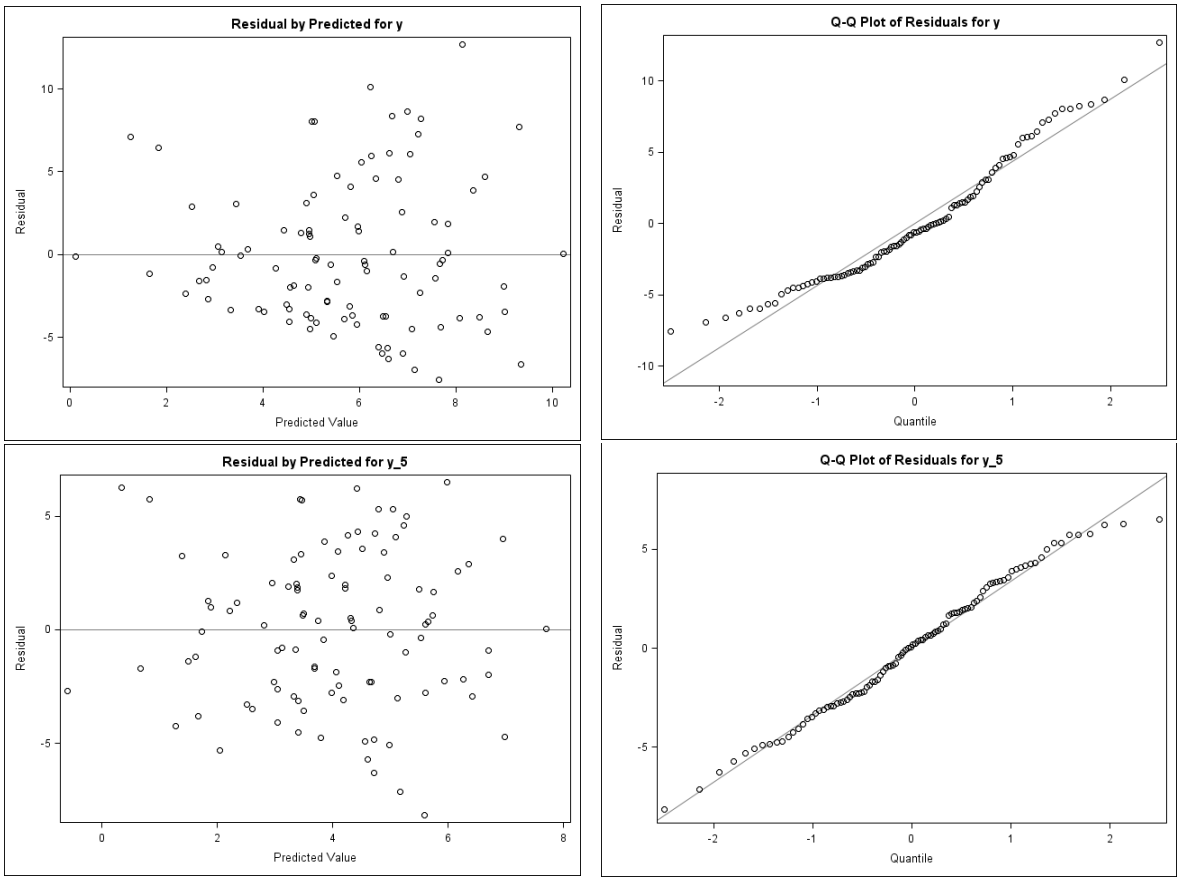
	Prin7	Prin8	Prin9	Prin10	Prin11
height	-.259751	0.018476	0.520184	-.036531	0.375786
weight	-.058566	0.003991	-.634360	-.174066	0.449619
bmi	0.168277	0.006166	0.554743	0.038138	0.069126
area	-.143725	-.017816	-.095960	0.175724	-.772725
age	0.148534	-.013550	-.001968	0.003505	-.002419
avtrel	-.397123	-.000850	-.061237	0.611781	0.146172
avtrsp	0.825059	0.015953	-.032894	0.221914	0.054851
int	-.081491	0.009784	0.071012	-.715838	-.172852
temp	-.055085	0.706224	-.008134	0.004308	-.015415
barm	-.090203	0.006020	0.003518	-.000885	-.001557
hum	-.042697	-.707081	0.009525	-.007438	0.015737

From the eigenanalysis of the correlation matrix, there appears to be other collinearity within the variables besides the intercept.

By looking at the elements of the 11th Principal Component, the elements corresponding to the height, weight, bmi, area, average treatmill elevation, average treadmill speed, the interaction, temperature, and humidity have values father from zero relative to age and barometric pressure.

2. Find the Box-Cox transformation of the simulated data (BoxCox.dat) and compare the residual plots of the raw and transformed data.

Box-Cox Transformation Information for y				
Lambda		R-Square	Log Like	
0.00		0.19	-147.31	
0.25		0.19	-124.38	*
0.50	+	0.18	-122.59	<
0.75		0.17	-131.76	
1.00		0.16	-147.64	
< - Best Lambda				
* - 95% Confidence Interval				
+ - Convenient Lambda				



The residual plot for the non-transformed y-values shows a slight pattern with the residuals being clustered closer together for smaller predicted values of y with the dispersion increasing as the predicted values of y increase. This indicates that the assumptions of homogeneity of variance and linearity might be violated. Due to curvature, the Q-Q plot for the non-transformed y-values also indicates that the assumption of Gaussian errors might be violated.

After transforming the y-values using a Box-Cox transformation with lambda=0.50, the residual plot is scattered equally regardless of the predicted value of y_5 and the Q-Q plot no longer contains a curve in the middle. This suggests the Box-Cox transformation of y does not violate the assumptions of homogeneity, linearity, nor Gaussian errors.

3. Investigators are interested in the effect of dermal nicotine exposure in a population of Latino tobacco workers in North Carolina. (Nicotine can be absorbed from tobacco leaves through the skin and can cause nicotine poisoning, which is characterized by nausea, vomiting, headache, and dizziness.) Data were collected on tobacco work tasks and risk factors for exposure to nicotine during a summer tobacco work season. Nicotine exposure was measured by levels of cotinine, a nicotine metabolite, contained in saliva. Other covariates of interest include age, body mass index, education, work conditions (working in wet conditions is believed to increase nicotine absorption), type of tobacco work ("priming" refers to picking or harvesting the tobacco and is expected to result in highest nicotine exposures, "barning" refers to putting the harvested tobacco into a barn for curing, "topping" refers to breaking the flower off the top of the plant, and "other" refers to farm work that does not involve tobacco contact, such as driving a truck), and smoking (smokers would also have nicotine exposure through cigarettes, and it is not known whether exposure to tobacco leaves would increase cotinine levels to a similar extent in both smokers and non-smokers).

   The variables are available in the file tobacco.dat and listed in the following order.
   - COTININE: salivary cotinine concentration (in ng/mL)
   - AGE: age (in years)
   - BMI: body mass index (in kg/m2)
   - EDUC: years of education
   - WET: takes value 1 if work conditions on day of measurement were wet and takes value 0 otherwise
   - TASK: takes value 1 for priming, 2 for barning, 3 for topping, and 4 for other work not involving tobacco contact
   - LNNSMOKE: natural logarithm of (1 + number of cigarettes smoked per day)

   To report a test, provide H_0, the test statistic, the degrees of freedom, the p-value, the decision (accept/reject H_0), and an interpretation of the result in terms of the subject matter.

- One-Way ANOVA: For these questions, use the log of salivary cotinine as the response and task as the only predictor.
  - Report a test of whether all cell means are equal.

    H_0: All cell-means equal ($\mu_{Priming} = \mu_{Barning} = \mu_{Topping} = \mu_{Other}$)
    H_1: At least one cell-mean differs from another cell-mean.

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Usual Overall Test	3	790.4453988	263.4817996	116.20	<.0001

    F (3, 690) = 116.20 with p < 0.0001.
    Reject H_0.
    There is evidence to suggest that at least two of the cell-means differ.

  - If your overall test of the task effect was significant, examine all pairwise comparisons using the Scheffe correction. Summarize your findings in a table including columns for the estimated mean difference, degrees of freedom, *F* statistic, *p*-value, and a Scheffe confidence interval for the mean difference. Explain your findings in language the investigator can understand.

Table 3.1: Pairwise Comparisons

	Estimated Mean Difference	Degrees of Freedom	F	p-value	Scheffe 95% Confidence Interval
Priming v. Barning	0.9208	1, 690	19.58	0.0002	[0.3376, 1.5040]
Priming v. Topping	1.6738	1, 690	131.87	<0.0001	[1.2654, 2.0823]
Priming v. Other	2.6993	1, 690	332.68	<0.0001	[2.2845, 3.1140]
Barning v. Topping	0.7531	1, 690	12.99	0.0049	[0.1675, 1.3386]
Barning v. Other	1.7785	1, 690	71.38	<0.0001	[1.1885, 2.3684]
Topping v. Other	1.0254	1, 690	47.26	<0.0001	[0.6074, 1.4434]

The average values for the log of salivary cotinine concentration (in mg/mL) found in Latino tobacco workers in North Carolina differs significantly depending on the task workers preformed.

- Provide a table of parameter estimates and standard errors using (a) cell mean coding and (b) reference cell coding, and give the interpretations of parameters in both coding schemes. In addition, provide the $\boldsymbol{C}$ and $\boldsymbol{\theta_0}$ matrices used to test the hypothesis that average cotinine levels for workers involved in priming are greater than the average continine levels for all other workers.

Table 3.2: Parameter Estimates

Cell Mean Coding			
	Estimates	Standard Errors	
Priming	4.5086	0.1022	Latino tobacco workers in NC whose task is priming have an average log of salivary cotinine concentration of 4.5086 mg/mL.
Barning	3.5878	0.1813	Latino tobacco workers in NC whose task is barning have an average log of salivary cotinine concentration of 3.5878 mg/mL.
Topping	2.8347	0.1039	Latino tobacco workers in NC whose task is topping have an average log of salivary cotinine concentration of 2.8347 mg/mL.
Other	1.8093	0.1070	Latino tobacco workers in NC whose task is not priming, barning, or topping have an average log of salivary cotinine concentration of 1.8093 mg/mL.
Reference Cell Coding – Solution 1			
	Estimates	Standard Errors	
Intercept	1.8093	0.1070	Latino tobacco workers in NC whose task is not priming, barning, or topping have an average log of salivary cotinine concentration of 1.8093 mg/mL.
Priming	2.6993	0.1480	Latino tobacco workers in NC whose task is priming have an average log of salivary cotinine concentration of 2.6993 mg/mL higher than those whose task is not priming, barning, or topping.
Barning	1.7785	0.2105	Latino tobacco workers in NC whose task is barning have an average log of salivary cotinine concentration of 1.7785 mg/mL higher than those whose task is not priming, barning, or topping.
Topping	1.0254	0.1492	Latino tobacco workers in NC whose task is topping have an average log of salivary cotinine concentration of 1.0254 mg/mL higher than those whose task is not priming, barning, or topping.
Reference Cell Coding – Solution 2			
	Estimates	Standard Errors	
Intercept	4.5086	0.1022	Latino tobacco workers in NC whose task is priming have an average log of salivary cotinine concentration of 4.5086 mg/mL.
Barning	-0.9208	0.2081	Latino tobacco workers in NC whose task is barning have an average log of salivary cotinine concentration of 0.9208 mg/mL lower than those whose task is priming.
Topping	-1.6738	0.1458	Latino tobacco workers in NC whose task is topping have an average log of salivary cotinine concentration of 1.6738 mg/mL lower than those whose task is priming.
Other	-2.6993	0.1480	Latino tobacco workers in NC whose task is not priming, barning, or topping have an average log of salivary cotinine concentration of 2.6993 mg/mL lower than those whose task is priming.

H_0: $\mu_{priming} = (\mu_{barning} + \mu_{topping} + \mu_{other})/3$

Cell Mean Coding: $H_0: \beta_{priming} - \frac{\beta_{barning}+\beta_{topping}+\beta_{other}}{3} = 0$

C = [1 −1/3 −1/3 −1/3] $\theta_0 = 0$

Reference Cell Coding – Solution 1:

$$H_0: \beta_0 + \beta_{priming} = \frac{(\beta_0)+(\beta_0+\beta_{barning})+(\beta_0+\beta_{topping})}{3} \equiv$$

$$H_0: \beta_0 + \beta_{priming} = \beta_0 + \frac{1}{3}(\beta_{barning} + \beta_{topping}) \equiv H_0: \beta_1 - \frac{1}{3}(\beta_2 + \beta_3) = 0$$

C = [0 1 −1/3 −1/3] $\theta_0 = 0$

Reference Cell Coding – Solution 2:

$$H_0: \beta_0 = \frac{(\beta_0+\beta_{other})+(\beta_0+\beta_{barning})+(\beta_0+\beta_{topping})}{3} \equiv$$

$$H_0: \beta_0 = \beta_0 + \frac{1}{3}(\beta_{barning} + \beta_{topping} + \beta_{other}) \equiv H_0: \frac{1}{3}(\beta_1 + \beta_2 + \beta_3) = 0$$

C = [0 1/3 1/3 1/3] $\theta_0 = 0$

Note all these solutions assume equal sample sizes in the groups. You could factor the group sample sizes into these calculations.

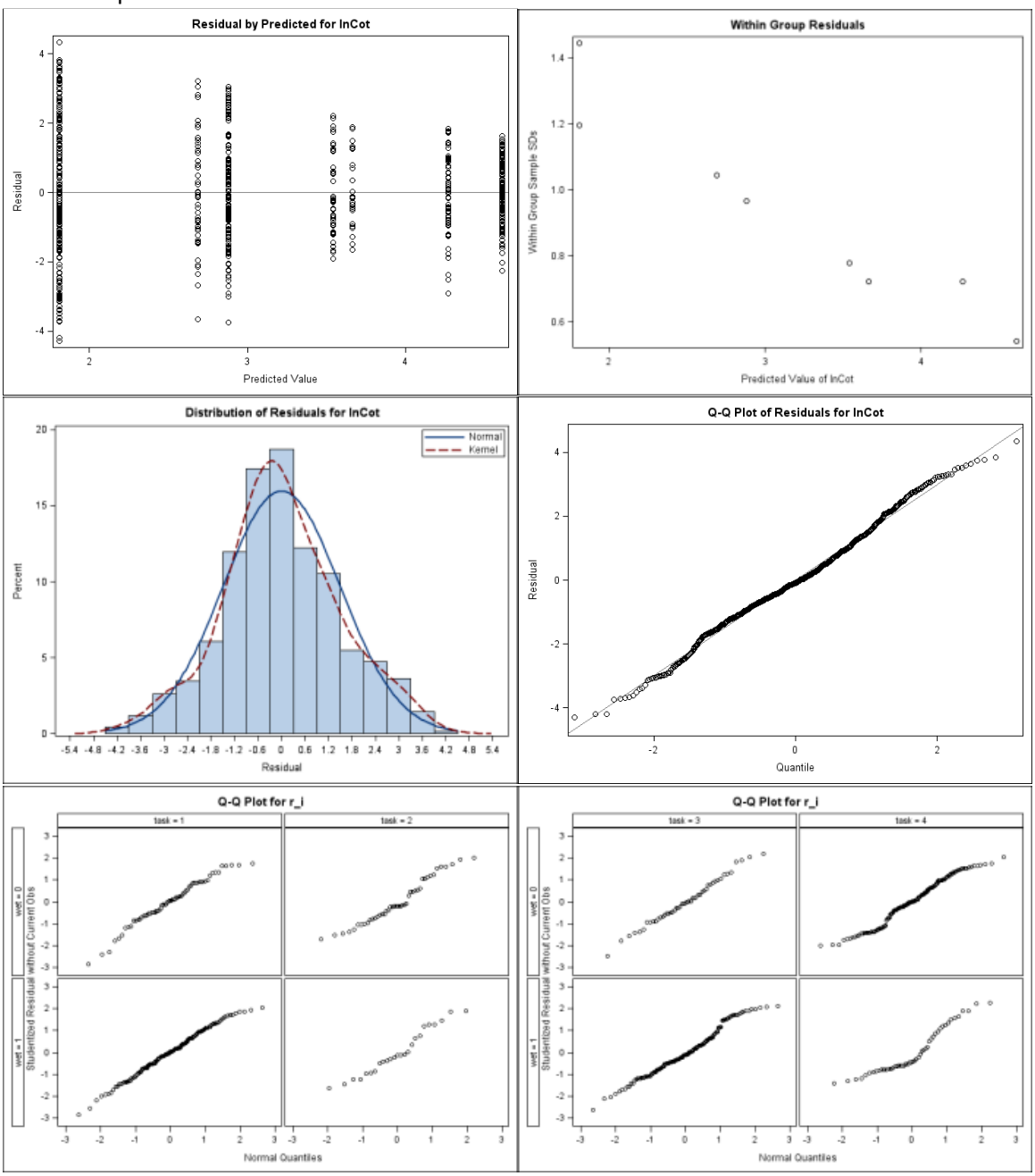
- Two-Way ANOVA: For these questions, use the log of salivary cotinine as the response and task and wet as predictors.
  - Fit the two-way ANOVA model with full interaction, and interpret all parameter estimates in your model, clearly stating which coding scheme you used. Discuss the validity of the HILE Gauss assumptions for this model.

Table of task by wet			
	wet		
	No	Yes	Total
task			
Priming	66	151	217
Barning	44	25	69
Topping	49	161	210
Other	148	50	198
Total	307	387	694

Complete & Not Balanced.

Cell Mean Coding - $\ln(cotinine) = w0t1\,\hat{\beta}_1 + w1t1\,\hat{\beta}_2 + w0t2\,\hat{\beta}_3 + w1t2\,\hat{\beta}_4 + w0t3\,\hat{\beta}_5 + w1t3\,\hat{\beta}_6 + w0t4\hat{\beta}_7 + w1t4\,\hat{\beta}_8$			
	Estimates	Standard Errors	
Priming & Not Wet	4.2693	0.1855	Latino tobacco workers in NC whose task is priming have an average log of salivary cotinine concentration of 4.2693 mg/mL in non-wet working conditions.
Priming & Wet	4.6131	0.1226	Latino tobacco workers in NC whose task is priming have an average log of salivary cotinine concentration of 4.6131 mg/mL in wet working conditions.
Barning & Not Wet	3.5427	0.2272	Latino tobacco workers in NC whose task is barning have an average log of salivary cotinine concentration of 3.5427 mg/mL in non-wet working conditions.
Barning & Wet	3.6670	0.3014	Latino tobacco workers in NC whose task is barning have an average log of salivary cotinine concentration of 3.6670 mg/mL in wet working conditions.
Topping & Not Wet	2.6882	0.2153	Latino tobacco workers in NC whose task is topping have an average log of salivary cotinine concentration of 2.6882 mg/mL in non-wet working conditions.
Topping & Wet	2.8793	0.1188	Latino tobacco workers in NC whose task is topping have an average log of salivary cotinine concentration of 2.8793 mg/mL in wet working conditions.
Other & Not Wet	1.8089	0.1239	Latino tobacco workers in NC whose task is not priming, barning, or topping have an average log of salivary cotinine concentration of 1.8089 mg/mL in non-wet working conditions.
Other & Wet	1.8105	0.2131	Latino tobacco workers in NC whose task is not priming, barning, or topping have an average log of salivary cotinine concentration of 1.8105 mg/mL in wet working conditions.

Assumptions:



Homogeneity – Within cells isn't as important as between cells. The scatter plot of the standard deviations of the residuals for each group reveals a potential pattern to the data. This assumption could be violated and since we have inequality of sample sizes between groups this could impact the testing accuracy.

Independence – given through the sample design.

Linearity – This is okay given the design.

Existence – finite sample satisfies this.

Gaussian Errors – The overall histogram of residuals and overall and individual QQ plots appear to support this assumption.

- *Based on this model*, create a table of the estimated mean log cotinine levels, associated standard errors, and how each estimated mean is obtained from the model parameters (e.g.,$\hat{\beta}_0 + \hat{\beta}_1$) for each task-wet combination.

Cell Mean Coding			
	Estimates	Standard Errors	$w0t1\,\hat{\beta}_1 + w1t1\,\hat{\beta}_2 + w0t2\,\hat{\beta}_3 + w1t2\,\hat{\beta}_4 + w0t3\,\hat{\beta}_5 + w1t3\,\hat{\beta}_6 + w0t4\hat{\beta}_7 + w1t4\,\hat{\beta}_8$
Grand Mean	3.1599	0.0699	$(\hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 + \hat{\beta}_4 + \hat{\beta}_5 + \hat{\beta}_6 + \hat{\beta}_7 + \hat{\beta}_8)/8$
Priming	4.4412	0.1112	$(\hat{\beta}_1 + \hat{\beta}_2)/2$
Barning	3.6049	0.1887	$(\hat{\beta}_3 + \hat{\beta}_4)/2$
Topping	2.7837	0.1229	$(\hat{\beta}_5 + \hat{\beta}_6)/2$
Other	1.8097	0.1232	$(\hat{\beta}_7 + \hat{\beta}_8)/2$
Wet	3.2425	0.1017	$(\hat{\beta}_2 + \hat{\beta}_4 + \hat{\beta}_6 + \hat{\beta}_8)/4$
Not Wet	3.0773	0.0961	$(\hat{\beta}_1 + \hat{\beta}_3 + \hat{\beta}_5 + \hat{\beta}_7)/4$

Cell Mean Coding			
	Estimates	Standard Errors	
Priming & Not Wet	4.2693	0.1855	$\hat{\beta}_1$
Priming & Wet	4.6131	0.1226	$\hat{\beta}_2$
Barning & Not Wet	3.5427	0.2272	$\hat{\beta}_3$
Barning & Wet	3.6670	0.3014	$\hat{\beta}_4$
Topping & Not Wet	2.6882	0.2153	$\hat{\beta}_5$
Topping & Wet	2.8793	0.1188	$\hat{\beta}_6$
Other & Not Wet	1.8089	0.1239	$\hat{\beta}_7$
Other & Wet	1.8105	0.2131	$\hat{\beta}_8$

- The Full Model in Every Cell: For these questions, use the log of salivary cotinine as the response and task, and lnnsmoke as predictors.
  - Fit the full model in every cell. Provide and interpret estimates of all parameters for this model.

Cell Mean Coding -

$$\ln(cotinine) = \hat{\beta}_{0t1}t1 + \hat{\beta}_{0t2}t2 + \hat{\beta}_{0t3}t3 + \hat{\beta}_{0t4}t4 + \hat{\beta}_{1t1}(t1lnnsmoke) + \hat{\beta}_{1t2}(t2lnnsmoke) + \hat{\beta}_{1t3}(t3lnnsmoke) + \hat{\beta}_{1t4}(t4lnnsmoke)$$

	Estimates	Standard Errors	
Priming	4.3344	0.0932	Latino tobacco workers in NC whose task is priming have an average log of salivary cotinine concentration of 4.3344 mg/mL.
Barning	3.1134	0.1666	Latino tobacco workers in NC whose task is barning have an average log of salivary cotinine concentration of 3.1134 mg/mL.
Topping	2.0123	0.0985	Latino tobacco workers in NC whose task is topping have an average log of salivary cotinine concentration of 2.0123 mg/mL.
Other	0.9137	0.0957	Latino tobacco workers in NC whose task is not priming, barning, or topping have an average log of salivary cotinine concentration of 0.9137 mg/mL.
Priming & LNNSMOKE	0.2946	0.0887	Latino tobacco workers in NC whose task is priming average log of salivary cotinine concentration increases by 0.2946 mg/mL for every 1 unit increase in lnnsmoke.
Barning & LNNSMOKE	0.7221	0.1450	Latino tobacco workers in NC whose task is barning average log of salivary cotinine concentration increases by 0.7221 mg/mL for every 1 unit increase in lnnsmoke.
Topping & LNNSMOKE	1.2305	0.0894	Latino tobacco workers in NC whose task is topping average log of salivary cotinine concentration increases by 1.2305 mg/mL for every 1 unit increase in lnnsmoke.
Other & LNNSMOKE	1.7789	0.1022	Latino tobacco workers in NC whose task is not priming, barning, or topping average log of salivary cotinine concentration increases by 1.7789 mg/mL for every 1 unit increase in lnnsmoke.

  - Report an appropriate test of whether task is related to cotinine levels. If this test is significant, report step-down tests to determine exactly where differences lie. For all tests reported, be sure to state H_0 clearly and give explicit justification for which tests were used and why.

Use a test of coincidence; the hypothesis indicates that the slopes and intercepts are all equal regardless of task.

H_0: $\beta_{0t1} = \beta_{0t2} = \beta_{0t3} = \beta_{0t4}$ and $\beta_{1t1} = \beta_{1t2} = \beta_{1t3} = \beta_{1t4}$

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
TEST OF COINCIDENCE	6	922.1700624	153.6950104	119.29	<.0001

$F(6, 687) = 119.29$ p-value < 0.001 Reject H_0

Task is related either by slope or intercept to cotinine levels at the 0.01 level.

Step down to determine if the differences are in the slopes or intercepts.

$H_0: \beta_{1t1} = \beta_{1t2} = \beta_{1t3} = \beta_{1t4}$

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
STEPDOWN: EQUAL SLOPES	3	169.6441128	56.5480376	43.89	<.0001

$F(3,690) = 43.89$ p-value < 0.0001 Reject H_0
The slopes are significantly different from each other at the 0.01 level.

$H_0: \beta_{0t1} = \beta_{0t2} = \beta_{0t3} = \beta_{0t4}$

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
STEPDOWN: EQUAL INTERCEPTS	3	902.0132306	300.6710769	233.36	<.0001

$F(3,690) = 233.36$ p-value < 0.0001 Reject H_0
The intercepts are significantly different from each other at the 0.01 level.

Use pair-wise tests to determine exactly which intercepts/slopes are different.

$\alpha = 0.05 / 6 = 0.0083$ $\sim F(1, 692)$

$H_0: \beta_{0ti} = \beta_{0tj}$

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
PAIRWISE INTERCEPTS T1 V T2	1	52.7322641	52.7322641	40.93	<.0001
PAIRWISE INTERCEPTS T1 V T3	1	377.7620263	377.7620263	293.20	<.0001
PAIRWISE INTERCEPTS T1 V T4	1	845.0460533	845.0460533	655.87	<.0001
PAIRWISE INTERCEPTS T2 V T3	1	41.7145812	41.7145812	32.38	<.0001
PAIRWISE INTERCEPTS T2 V T4	1	168.9622476	168.9622476	131.14	<.0001
PAIRWISE INTERCEPTS T3 V T4	1	82.4489524	82.4489524	63.99	<.0001

Reject H_0 for all ij pairs
All intercepts are significantly different from each other at the 0.01 level.

$H_0: \beta_{1ti} = \beta_{1tj}$

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
PAIRWISE SLOPES T1 V T2	1	8.1529052	8.1529052	6.33	0.0121
PAIRWISE SLOPES T1 V T3	1	71.1683903	71.1683903	55.24	<.0001
PAIRWISE SLOPES T1 V T4	1	155.0229026	155.0229026	120.32	<.0001
PAIRWISE SLOPES T2 V T3	1	11.4843308	11.4843308	8.91	0.0029
PAIRWISE SLOPES T2 V T4	1	45.7480465	45.7480465	35.51	<.0001
PAIRWISE SLOPES T3 V T4	1	21.0166241	21.0166241	16.31	<.0001

Reject H_0 for all ij pairs except i=1 and j=2
Except for task 1 and 2, the slopes are significantly different from each other at the 0.01 level.

BIOS663 Midterm Spring 2013
Thursday, March 7, 2013

Instructions: Please be as rigorous as possible in all of your answers and show all your work.

Please sign the honor code pledge and submit it with your report. Violation of the honor code below will be prosecuted (penalties may include failure of the course and expulsion from the university).

Honor Code Pledge: On my honor, I have neither given nor received aid on this examination.

Name:

Signature:

Date:

1. *(20 points total)* MULTIPLE CHOICE QUESTIONS (Please circle the best answer).

   - *(5 points)* Which choice is not an appropriate description of $\hat{y}$ in a regression model?
     A. Estimated response
     B. Predicted response
     C. Estimated average response
     D. Observed response
     Solution: D
   - *(5 points)* Which of the following is the best way to determine whether or not there is a statistically significant linear relationship between two variables?
     A. Compute a regression line from a sample and see if the sample slope is 0.
     B. Compute the correlation coefficient and see if it is greater than 0.5 or less than 0.5.
     C. Conduct a test of the null hypothesis that the population slope is 0.
     D. Conduct a test of the null hypothesis that the population intercept is 0.
     Solution: C
   - *(5 points)* Which of the following case diagnostic measures is based on Y values only (and not X values)?
     A. Cooks distance
     B. Studentized residual
     C. Leverage
     D. None of the above
     Solution: D
   - *(5 points)* Which of the following is NOT true for the linear regression model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i (i = 1, \cdots, 100)$ where all 5 model assumptions hold.
     A. $\hat{\beta}_1 + \beta_2$ is a statistic
     B. $\hat{y}_i$ can be uniquely predicted from the above model
     C. β_1 may not be always estimable
     D. the residuals from the model are summed to 0
     Solution: A

2. *(40 points total)* You are working on a statistical consulting lab. One day, a client came with a gas consumption data. In this study, the client is interested in modeling the fuel efficiency of automobiles. A typical measure of fuel efficiency used by EPA and car manufactures is "gallons/100 miles". The client collected data on 100 cars. He measured two explanatory variables, x1=weight (in unit of 1000lb); and x2=number of cylinders. He also measured the fuel efficiency of each car (in "gallons/100 miles"). Let $\mathbf{X} = (\mathbf{J}_n, \mathbf{x}_1, \mathbf{x}_2)$ and the linear regression model considered is

$$y = \beta_0 + \beta_1 x1 + \beta_2 x2 + error.$$

Potentially helpful results:
$(X'X)^{-1} = \begin{bmatrix} 0.308 & -0.06 & -0.017 \\ -0.06 & 0.025 & -0.004 \\ -0.017 & -0.004 & 0.006 \end{bmatrix}$ and $X'y = \begin{bmatrix} 405 \\ 1402 \\ 2350 \end{bmatrix}$.

(a) *(8 points)* A partial ANOVA table for testing the association of the three covariates with the response y is given below. Complete the table.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	???	79	???	???	<0.001
Error	???	11	???		
Corrected Total	???	???			

Solution:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	79	39.5	349.6	<0.001
Error	97	11	0.113		
Corrected Total	99	90			

(b) *(6 points)* Fill in the cells with ??? in the following table.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	???	???	???	0.009
x1	???	???	???	<0.001
x2	???	???	???	<0.001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	0.67	0.187	3.58	0.009
x1	1.35	0.053	25.47	<0.001
x2	1.61	0.026	61.81	<0.001

(c) *(6 points)* Test the following hypothesis: $H_0 : \beta_1 = 1$.
Solution:
$t - test = \frac{\hat{\beta}_1 - 1}{SE(\hat{\beta}_1)} = \frac{1.35-1}{0.053} = 6.6 \sim t_{97}$ which is greater than 1.96, so reject the null hypothesis at $\alpha = 0.05$.

(d) *(6 points)* Test the following hypothesis: $H_0 : \beta_1 = \beta_2 = 1$.
Solution:
Let $\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ and $\boldsymbol{\theta}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, then $H_0 : C\boldsymbol{\theta} = \boldsymbol{\theta}_0$.

So $M = \mathbf{C}(X'X)^{-1}\mathbf{C}' = \begin{bmatrix} 0.025 & -0.004 \\ -0.004 & 0.006 \end{bmatrix}$ with $M^{-1} = \begin{bmatrix} 44.78 & 29.85 \\ 29.85 & 186.57 \end{bmatrix}$ and $\hat{\boldsymbol{\theta}} = (1.35, 1.61)'$.

Thus

$F - test = \frac{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)' M^{-1} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/2}{\hat{\sigma}^2} = \frac{43.83}{0.113} = 387.8 \sim F_{2,97}$

(e) *(6 points)* Find a 95% confidence interval of $\beta_1 + \beta_2$.

Solution:

Let $\theta = \beta_1 + \beta_2$ then $\hat{\theta} = \hat{\beta}_1 + \hat{\beta}_2 = 2.96$.

$\hat{var}(\hat{\theta}) = (0, 1, 1)\hat{var}(\hat{\boldsymbol{\beta}})(0, 1, 1)' = 0.0026$ and $SE(\hat{\theta}) = \sqrt{0.0026} = 0.051$

95 % CI of θ is $2.96 \pm 1.96 * 0.051 = [2.86, 3.06]$

(f) *(8 points)* Now you decide to transform x1 and x2 to z1=x1 - 2 and z2=x2-4 where 2 and 4 refer the population minimal car weight and minimal number of cylinders. Refit data with the following linear model $y = \beta_0^* + \beta_1^* z1 + \beta_2^* z2 + error$. Please describe the meaning of β_0^* and fill in the following table:

Parameter	Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	???	???	???	-
z1	???	???	???	???
z2	???	???	???	???

Parameter	Estimate	Standard Error	t Value	Pr > |t|
(Intercept)	9.8	0.085	115.3	-
21	1.35	0.053	25.47	<0.001
22	1.61	0.026	61.81	<0.001

3. *(40 points total)* Consider the set of hypothetical data below $\mathbf{y}_{5\times1} = \mathbf{X}_{5\times3}\boldsymbol{\beta}_{3\times1} + \boldsymbol{\varepsilon}_{5\times1}$, where

$$\mathbf{y} = \begin{bmatrix} 0 \\ 2 \\ 4 \\ 6 \\ 10 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 6 & 11 \\ 1 & 7 & 13 \\ 1 & 8 & 15 \\ 1 & 9 & 17 \\ 1 & 11 & 21 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

   (a) *(5 points)* Is there a problem of multicollinearity in this regression? Prove or disprove that there exists no multicollinearity problem.

   Solution: yes since the rank of the desing matrix is 2 instead of 3.

   (b) *(5 points)* Can you compute OLS estimates of the three parameters and explain why.

   Solution: No since the design matrix is not full rank.

   (c) *(5 points)* Throwing out any redundant columns of the X matrix if necessary and re-express the model as $\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ where $\mathbf{X}^*$ is full rank. Express $\boldsymbol{\beta}^*$ in terms of $\boldsymbol{\beta}$.

   One solution is to let $\mathbf{X}^* = \begin{bmatrix} 1 & 6 \\ 1 & 7 \\ 1 & 8 \\ 1 & 9 \\ 1 & 11 \end{bmatrix}$ and $\boldsymbol{\beta}^* = \begin{bmatrix} \beta_0 - \beta_2 \\ \beta_1 + 2\beta_2 \end{bmatrix}$.

   (d) *(5 points)* Suppose that there are two students in the Bios663 class whose names are Jim and Chris. Suppose further that they estimated the parameters β_0, β_1 and β_2 by trial and error. As a result, they got different answers, i.e., (-6, -10, 6) and (-10, -2, 2) respectively. And each of them argues that his answer is better. What do you think about these two answers? Which answer fits better to the data?

   Solution: the two solutions are the same since they give the same estimated regression model.

   (e) *(8 points)* Compute a 95% confidence interval for the mean response of individuals with $x_1 = 1$ and $x_2 = 1$. Do you think the model provides a good estimate for this mean response? Why?

   Solution: SSE from the model is 0 and also we can check that the mean respose for $x_1 = 1$ and $x_2 = 1$ is estimable, so the 95% CI is $[-10, -10]$. The model does not provide a good estimate for this mean response since $x_1 = 1$ is far outside the range of the observed x_1 values.

   (f) *(6 points)* Show as rigorously as possible whether $H_0 : \beta_0 - \beta_2 = 0$ &$\beta_1 + 2\beta_2 = 2$ &$2\beta_0 + \beta_1 = 2$ is testable. If not, can it be reduced to an equivalent testable hypothesis? If yes, present an equivalent testable hypothesis.

Solution: it is not testable but can be reduced to a testable hypothesis, such as $H_0 : \beta_0 - \beta_2 = 0 \ \&\beta_1 + 2\beta_2 = 2$.

(g) *(6 points)* Show as rigorously as possible whether $H_0 : \beta_0 + \beta_1 = 0$ is testable. If so, report your test.

Solution: it is not testable.

1. (10 pts) For a general linear regression problem with p covariates (including intercept and $p-1$ additional covariates) and sample size n, the regression model can be written as $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$, where $\mathbf{e} \sim N(0, \sigma^2 I_{n\times n})$, and $\mathbf{X}$ is full rank.

   (a) (4 pts) What are the dimensions of matrix/vector of $\mathbf{y}$, $\mathbf{X}$, β, and $\mathbf{e}$? What is the rank of $\mathbf{X}$? Please explain why $\text{cov}(\mathbf{e}) = \sigma^2 I_{n\times n}$ implies the assumptions of independence and homogeneity.

$y_{n\times 1}$ $X_{n\times p}$ $e_{n\times 1}$ rank $(X)=p$ diagonals

$Cov(e) = \sigma^2 \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} \Rightarrow$ { independence since off diagonals are 0
homogeneity since $Var(y_i)=\sigma^2$ for all i

   (b) (4 pts) Derive the least squares estimates: $\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{y})$ by minimizing the least squares objective function, i.e., to minimize $(\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{y}-\mathbf{X}\beta)$.

See lecture note

   (c) (4pts) Calculate $E(\hat{\beta})$ and $\text{cov}(\hat{\beta})$.

See lecture note

2. (8pts) Now assume $\mathbf{e} \sim N(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \text{cov}(\mathbf{e})$ and $\boldsymbol{\Sigma}$ is positive definite. Given the eigen-value decomposition of $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Gamma}\mathbf{V}'$, where $\boldsymbol{\Gamma}$ is a diagonal matrix and $\mathbf{V}$ is an orthonormal matrix such as $\mathbf{V}\mathbf{V}' = \mathbf{V}'\mathbf{V} = I_{n\times n}$, we define $\boldsymbol{\Sigma}^{-1/2} = \mathbf{V}\boldsymbol{\Gamma}^{-1/2}\mathbf{V}'$. Let $\tilde{\mathbf{y}} = \boldsymbol{\Sigma}^{-1/2}\mathbf{y}$, $\tilde{\mathbf{X}} = \boldsymbol{\Sigma}^{-1/2}\mathbf{X}$, and $\tilde{\mathbf{e}} = \boldsymbol{\Sigma}^{-1/2}\mathbf{e}$. We consider a linear regression problem $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\alpha} + \tilde{\mathbf{e}}$.

(a) (2 pts) Please show $\text{cov}(\tilde{\mathbf{e}}) = I_{n\times n}$.

by matrix multiplication for diagonal matrix

$$\text{cov}(\tilde{e}) = \Sigma^{-\frac{1}{2}}\Sigma\Sigma^{-\frac{1}{2}}$$
$$= V\Gamma^{-\frac{1}{2}}V'V\Gamma V'V\Gamma^{-\frac{1}{2}}V'$$
$$= V\Gamma^{-\frac{1}{2}}\Gamma\Gamma^{\frac{1}{2}}V' = VV' = I$$

(b) (2 pts) For a linear regression problem $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, the formula for least squares estimates is: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}(\mathbf{X}^\intercal\mathbf{y})$. Please use this formula to calculate the least squares estimates of regression coefficients $\hat{\boldsymbol{\alpha}}$ for the regression model $\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\alpha} + \tilde{\mathbf{e}}$, in terms of $\mathbf{X}$, $\mathbf{y}$ and $\boldsymbol{\Sigma}$.

$$\hat{\alpha} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{Y}$$
$$= (X^T\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}X)^{-1}X^T\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}}Y$$
$$= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$$

(c) (4pts) Calculate $E(\hat{\boldsymbol{\alpha}})$ and $\text{cov}(\hat{\boldsymbol{\alpha}})$.

$$E(\hat{\alpha}) = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}X\beta = \beta$$

$$\text{Cov}(\hat{\alpha}) = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\Sigma\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}$$
$$= (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}$$
$$= (X^T\Sigma^{-1}X)^{-1}$$

3. (20pts) We are interested in data collected by the Environmental Protection Agency (EPA) at the Health Effects Research Laboratory at UNC: Chapel Hill. One hundred seventy young adult males received a battery of pulmonary function tests. Fit a model with average forced vital capacity (FVC) (in ml) as the outcome and height, weight, body mass index (BMI=$\frac{\text{weight (kg)}}{(\text{height (m)})^2}$), body surface area, age, average treadmill elevation, average treadmill speed, temperature, barometric pressure, and humidity as predictors.

   (a) (5pts) To assess for possible co-linearity in the covariates, we perform PCA on the correlation matrix of this data. As shown in the following output, the 10-th eigen-value is very small, which means a particular

Eigenvalue decomposition of the Correlation Matrix

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.98457157	0.95976513	0.2985	0.2985
2	2.02480644	0.36097943	0.2025	0.5009
3	1.66382702	0.63279205	0.1664	0.6673
4	1.03103496	0.06376972	0.1031	0.7704
5	0.96726525	0.20929379	0.0967	0.8672
6	0.75797146	0.20748316	0.0758	0.9429
7	0.55048830	0.53278371	0.0550	0.9980
8	0.01770459	0.01584173	0.0018	0.9998
9	0.00186286	0.00139531	0.0002	1.0000
10	0.00046755		0.0000	1.0000

Eigenvectors

		Prin1	Prin2	Prin3	Prin4
height	Height (cm)	0.429342	0.036906	0.384653	-.240983
weight	Weight (kg)	0.562292	-.092679	-.095943	0.026718
bmi		0.340930	-.147689	-.443854	0.239531
area	Body Surface Area (M**2)	0.566969	-.047500	0.092208	-.079656
age	Age (years)	0.084799	-.102998	-.198442	-.321636
avtrel	Average Treadmill Elevation (deg)	-.116240	-.026373	0.497176	0.182497
avtrsp	Average Speed of Treadmill (mph)	0.144346	0.094273	0.571216	0.156073
temp	Air Temperature (deg C)	0.084996	0.675223	-.123262	0.099538
barm	Barometric Pressure (mmHg)	0.070861	-.175834	-.013681	0.832373
hum	Relative Humidity %	0.089569	0.677455	-.095377	0.116735

Eigenvectors

	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10
height	-.120483	-.361837	0.222180	0.018428	0.521355	0.375921
weight	-.012319	0.158716	0.071437	0.004011	-.639418	0.475397
bmi	0.092604	0.520638	-.117929	0.006137	0.557078	0.060451
area	-.061800	-.035176	0.137407	-.017784	-.093401	-.792758
age	0.856166	-.289687	-.149123	-.013509	-.001675	-.003181
avtrel	0.442067	0.455127	0.550162	0.007661	-.000498	-.001474
avtrsp	0.112098	0.166298	-.760874	0.019013	-.010800	0.001191
temp	0.096489	-.017527	0.055784	0.706187	-.007909	-.015994
barm	0.121960	-.502455	0.060375	0.005988	0.003385	-.001295
hum	0.087996	-.009950	0.046292	-.707073	0.009011	0.017050

linear combination of the covariates has small variance. Which linear combination it is? Explain why is it possible that this combination has small variance? Could this PCA captures co-linearity between intercept and other covariates? and why?

approximately .4 height + .5 weight — .8 area

No intercept effect has been removed from correlation matrix since it remove mean values

(b) (4pts) Consider a linear regression model with all the covariates. Let $\beta = (\beta_0, \beta_1, ..., \beta_{10})^T$ be the intercept and the regression coefficients for height, weight, bmi, area, age, avtrel, avtrsp, temp, barm, and hum, respectively. Test the hypothesis: H_0: $\beta_1 = \beta_2 = 2\beta_4$ using general linear hypothesis. Please write down C and θ_0 so that the test can be written $C\beta = \theta_0$, and please write down the formula of test-statistic while denoting the data matrix for intercept and the 10 covariates by $\mathbf{X}$, and denoting the residual variance of this linear regression model by $\hat{\sigma}^2$.

$$C = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & -2 & \cdots & 0 \end{pmatrix} \qquad \theta_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$F = \frac{(\hat{\theta} - \theta_0)' M^{-1} (\hat{\theta} - \theta_0)/2}{\hat{\sigma}^2}$$

$$M = C(X'X)^{-1}C'$$

(c) (7pts) After a few rounds of testing, we decide to have final model without area, temp, hum, and barm.

i. (2pts) Based on the following ANOVA table, what is the R^2? Please show your calculation and you may round those numbers to simplify the calculation.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	49984918	8330820	20.81	<.0001
Error	163	65242013	400258		
Corrected Total	169	115226931			

Root MSE	632.65927	R-Square	0.4130
Dependent Mean	5335.43235	Adj R-Sq	
Coeff Var	11.85769		

$R^2 \approx \frac{50}{115}$

ii. (2pts) Based on the following t-table, if we test whether the regression coefficient for age is 0 by added last test, what is the value of F-statistic, and what are the degrees of freedom?

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	4899.55068	13356	0.37	0.7142
height	Height (cm)	1	-32.70393	75.89989	-0.43	0.6671
weight	Weight (kg)	1	119.42970	92.38958	1.29	0.1980
bmi		1	-286.38260	297.80536	-0.96	0.3377
age	Age (years)	1	27.86381	13.61020	2.05	0.0422
avtrel	Average Treadmill Elevation (deg)	1	51.50263	37.65313	1.37	0.1733
avtrsp	Average Speed of Treadmill (mph)	1	755.16631	379.56118	1.99	0.0483

$F = (2.05)^2$

$df = (1, 163)$

iii. (3pts) Based on this reduced model with 6 covariates, which characteristics are associated with the best (largest) FVC?

Shorter, heavier, low bmi, older, higher avtrel and higher avtrsp

(d) (4pts) In the diagnosis of this model, we detect a few data points as outliers based on either leverage or cook's distance. Please explain what are the difference of leverage and cook's distance.

high leverage means outlier in X

large Cook's distance means high influence on regression

4. (20pts) Consider a linear regression problem to study the association between the physical activity of 12 mice vs. environment (0 for standard environment and 1 for enriched one) and dosage of a drug (with dosage 0, 1, and 2).

observation	activity	environment	drug
1	102	0	0
2	97	0	0
3	102	0	1
4	82	0	1
5	108	0	2
6	111	0	2
7	95	1	0
8	100	1	0
9	106	1	1
10	110	1	1
11	118	1	2
12	116	1	2

(a) (4pts) First consider a linear model with two covaraites:

$$E(\texttt{activity}) = b_0 + b_1\texttt{environment} + b_2\texttt{dose}$$

If we write the above model by a matrix form: $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, what are the meanings of $\mathbf{y}$, $\mathbf{X}$ and $\mathbf{e}$, and what are their dimensions?

$y_{12\times 1}$ $X_{12\times 3}$ $e_{12\times 1}$

response covariate error

(b) (4pts) Please calculate the correlation between two variables: environment and drug. For added in-order test, would the p-values for environment and drug remain the same for two orders: environment followed by drug; and drug followed by environment?

$$Cor(en, drug) = \frac{E(XY) - E(X)E(Y)}{\uparrow \quad \uparrow} = \frac{\sum x_i y_i}{12} - \frac{\sum x_i}{12}\frac{\sum y_i}{12} = 0$$

(c) (8pts) Given the following regression coefficient estimates and type III ANOVA table.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	92.958	4.000	23.240	2.41e-09
enviorment	7.167	4.276	1.676	0.1281
drug	7.375	2.619	2.816	0.0202

Response: activity

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
enviorment	1	154.08	154.08	2.8088	0.12806
drug	1	435.12	435.12	7.9321	0.02017
Residuals	9	493.71	54.86		

Please test the null hypothesis $H_0 : b_1 = b_2 = 0$ using (1) general linear hypothesis testing and (2) comparison of the sum squares of two models. Write down your test statistic, its asymptotic distribution and the degree of freedom. You should plug in the numbers into your formula of test statistic but do not need to calculate it. If you need $(\mathbf{X}'\mathbf{X})^{-1}$. Simply use $(\mathbf{X}'\mathbf{X})^{-1}$ rather than the actual numbers.

$\theta = C b$

(1) GLH $\quad C = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \theta_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

$$F = \frac{(\hat{\theta} - \theta_0)\, \Lambda^{-1} (\hat{\theta} - \theta_0)/2}{\hat{\sigma}^2} \qquad df = (2, 9)$$

$$\Lambda = C(X'X)^{-1}C'$$

(2) Model 1 $\quad y \sim \beta_0$

Model 2 $\quad y \sim \beta_0 + \beta_1 en + \beta_2 drug$

$$F = \frac{(SS_2 - SS_1)/2}{MSE_2} = \frac{(154.08 + 435.12)/2}{493.71/9} \qquad df = (2, 9)$$

Model $y \sim \beta_0 + \beta_1 \text{drug}$ $R_1^2 = \frac{453}{154+453+494}$

Model 2 $y \sim \beta_0 + \beta_1 \text{en} + \beta_2 \text{drug}$ $R_2^2 = \frac{453+154}{154+453+494}$

(d) (4pts) What is the R^2 of a smaller model with intercept and drug? What is the R^2 of a larger model with intercept, environment, and drug? Feel free to use approximations in your calculation. Then if we double the sample size from 12 to 24, while assuming the R^2 of these two models remain the same, what would be the F-statistic to test the null hypothesis that the regression coefficient for environment is 0.

$$F = \frac{(SSS_2 - SSS_1)/1}{SSE_2/(n-p)}$$

SSY = sum squares of y

$$= \frac{(SSS_2 - SSS_1)}{(SSY - SSS_2)(n-p)}$$

$$= \frac{R_2^2 - R_1^2}{(1 - R_2^2)/(n-p)}$$

$$\frac{F_{new}}{F_{old}} = \frac{n_{new} - p}{n_{old} - p} = \frac{24-3}{12-3} = \frac{21}{9}$$

BIOS663 Midterm Exam Spring 2019
March 6, 2019.

Instructions: Please be as rigorous as possible in all of your answers and show all your work.

Please sign the honor code pledge and submit it with your report. Violation of the honor code below will be prosecuted (penalties may include failure of the course and expulsion from the university).

Honor Code Pledge: On my honor, I have neither given nor received aid on this examination.

Name:

Signature:

Date:

1. *(20 points total)* Suppose $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \sim N(0, \mathbf{\Sigma})$ where $\mathbf{\Sigma} = \begin{bmatrix} 1 & 0 & 0.6 \\ 0 & 1 & 0.5 \\ 0.6 & 0.5 & 1 \end{bmatrix}$

   - (7 points) Derive the distribution of $2x_1 + x_2 - x_3$.
     Solution: Let $c = (2, 1, -1)$, then $var(2x_1+x_2-x_3) = c\mathbf{\Sigma}c^{'} = 2.6$ thus $2x_1+x_2-x_3$ follows a normal distribution with mean 0 and variance 2.6.
   - (7 points) Calculate $Cov(x_1 - x_2, 2x_2 + x_3)$.
     Solution: Let $c1 = (1, -1, 0)$ and $c2 = (0, 2, 1)$, then $Cov(x_1 - x_2, 2x_2 + x_3) = c1\mathbf{\Sigma}c2^{'} = -1.9$.
   - (6 points) Prove or dis-prove (with details) that $\mathbf{A} = \begin{bmatrix} 2 & 1 & 3 \\ 1 & 0 & 2 \\ 4 & 1 & -2 \end{bmatrix}$
     has linearly independent columns.
     Solution: Since $\|\mathbf{A}\| = 9 \neq 0$, the rank of $\mathbf{A}$ is thus full rank, and $\mathbf{A}$ has linearly independent columns.

2. *(40 points total)* Consider the model $\mathbf{y} = \beta_0\mathbf{1} + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \boldsymbol{\varepsilon}$, where

$$\mathbf{y} = \begin{bmatrix} 2 \\ 1 \\ 3 \\ 5 \\ 6 \end{bmatrix}, \quad \mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_1 = \begin{bmatrix} -2 \\ -1 \\ 1 \\ 0 \\ 3 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} -1 \\ -1 \\ 3 \\ -1 \\ -2 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

   with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I})$. Potentially helpful facts:
   A: the corrected *total* sum of squares is 17.2,
   B: a generalized inverse (if $\mathbf{X}$ is full rank, this inverse is unique) of $\mathbf{X}'\mathbf{X}$ is
   $(\mathbf{X}'\mathbf{X})^{-} = \begin{bmatrix} 0.21 & -0.02 & 0.02 \\ -0.02 & 0.07 & -0.02 \\ 0.02 & -0.02 & 0.08 \end{bmatrix}$; $\mathbf{X}'y = \begin{bmatrix} 17 \\ 16 \\ -5 \end{bmatrix}$ and
   C: 97.5 percentiles of student t-distributions:

df	1	2	3	4	5
	12.706	4.303	3.182	2.776	2.571

   - *(8 points)* Compute the least square estimates of the model parameters and their standard errors.

     Solution: $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y = (3.15, 0.88, -0.38)^{'}$; $\hat{y} = \mathbf{X}\hat{\beta} = (1.77, 2.65, 2.89, 3.53, 6.17)'$ and $\hat{\epsilon} = y - \hat{y} = (0.23, -1.65, 0.11, 1.47, -0.17)^{'}$. Thus $\hat{\sigma}^2 = 2.49$ and $\hat{Var}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$, leading to $se(\hat{\beta}_0) = \sqrt{0.52} = 0.72, se(\hat{\beta}_1) = \sqrt{0.17} = 0.41$, and $se(\hat{\beta}_2) = \sqrt{0.2} = 0.45$.

- *(8 points)* Compute the 95% prediction interval for a subject with $x_1 = 1$ and $x_2 = 2$.

  Solution: $\hat{y} = (1, 1, 2)\hat{\beta} = 3.27$ and $var(\hat{y}) = (1, 1, 2)\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}(1, 1, 2)' + \hat{\sigma}^2 = 3.9$ a 95% prediction interval is $3.27 \pm 4.3\sqrt{3.9} = (-5.3, 11.7)$.

- *(8 points)* Calculate the corrected ${R^2}_c$, interpret its value, and test the hypothesis that its corresponding population value is zero, that is, $H_0 : \rho_c^2 = 0$.

  Solution: Since $\bar{y} = 3.4$, we have $CSS(regression) = \sum_i \hat{y}_i^2 - 5\bar{y}^2 = 11.2$ $CSS(total) = \sum_i y_i^2 - 5\bar{y}^2 = 17.2$ and corrected ${R^2}_c = 11.2/17.2 = 0.65$.
  $H_0 : \rho_c^2 = 0$ is equivalent to $H_0 : \beta_1 = \beta_2 = 0$. Thus we have $F - test = \frac{11.2/2}{\hat{\sigma}^2} = 2.25 \sim F_{2,2}$.

- *(6 points)* Consider the following hypothesis test: $E(y \mid \text{covariates of individual 5}) = 2E(y \mid \text{covariates of individual 1})$. Give $\mathbf{C}$, $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_0$ that are associated with the hypothesis test. Show as rigorously as possible whether your $\boldsymbol{\theta}$ is testable. If so, test the hypothesis.

  Solution: $\beta_0 + 3\beta_1 - 2\beta_2 = 2(\beta_0 - 2\beta_1 - \beta_2)$ which is $\beta_0 - 7\beta_1 = 0$. Let $\mathbf{C} = (1, -7, 0)$, then $\theta = \beta_0 - 7\beta_1$. For $H_0 : \theta = 0$, the associated t-test $= \hat{\theta}/se(\hat{\theta}) = -3.01/3.124 = -0.96$. Since $\| - 0.96\| < 4.3$, the hypothesis is not statistically significant given type I error of 0.05.

- *(5 points)* Show as rigorously as possible whether $\begin{pmatrix} \beta_0 + \beta_1 \\ \beta_1 \\ \beta_0 + 2\beta_1 - 3\beta_2 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ 2\beta_2 \\ 2 \end{pmatrix}$ is testable. If so, test the hypothesis. If not, can you construct an equivalent test that is testable? If yes, perform the equivalent test. If not, explain why.

  Solution: Let $\mathbf{C} = \begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & -2 \\ 1 & 2 & -3 \end{bmatrix}$, then $\boldsymbol{\theta} = \mathbf{C}\boldsymbol{\beta}$ which is estimable since $\mathbf{X}$ is full rank.
  The test is $H_0 : \boldsymbol{\theta} = \begin{bmatrix} 0 \\ 0 \\ 2 \end{bmatrix}$. Since $\mathbf{C}$ is not full rank, the test is not testable. Further, the test cannot be reduced to a testable hypothesis, since the three equations conflict with each other.

- *(5 points)* Show as rigorously as possible whether $\begin{pmatrix} \beta_0 + \beta_1 \\ \beta_1 \\ \beta_0 + 2\beta_1 - 3\beta_2 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ 2\beta_2 \\ 0 \end{pmatrix}$ is testable. If so, test the hypothesis. If not, can you construct an equivalent test that is testable? If yes, perform the equivalent test. If not, explain why.

Solution: Follow the above question, with the same $\mathbf{C}$ and , the test now is $H_0 : \boldsymbol{\theta} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$ which again is not testable. However, we can reduce the above test to $\begin{pmatrix} \beta_0 + \beta_1 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} \beta_2 \\ 2\beta_2 \end{pmatrix}$ or $\begin{pmatrix} \beta_0 + \beta_1 - \beta_2 \\ \beta_1 - 2\beta_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ with $\mathbf{C} = \begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & -2 \end{bmatrix}$ and $\hat{\boldsymbol{\theta}} = (4.41, 1.64)'$

F-test $= \frac{\hat{\boldsymbol{\beta}}'(\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1})\hat{\boldsymbol{\theta}}/2}{\hat{\sigma}^2} = 34.2/2.49 = 13.7 \sim F_{2,2}$.

3. *(40 points total)* An investigator at UNC conducted a survey of Chapel Hill residents both before and after construction of a new exercise trail. Before the trail was constructed, she determined the baseline physical activity levels of a number of Chapel Hill residents. After construction of the trail, she interviewed the same group of residents about their physical activity levels (after construction of the trail) along with their gender and age.

   Short descriptions of the variables of interest are provided below.

   - post: Average physical activity, measured in hrs per day, after construction of the trail.
   - pre: Physical activity, measured in hrs per day, before construction of the trail (baseline).
   - age: Age of each participant.
   - gender: Gender of each participant (Male =0 and Female=1).

   The investigator fit the following model, with data centered as indicated, to the physical activity data: $post = \beta_0 + \beta_1 pre + \beta_2 age + \beta_3 gender + error$. Let the design matrix of the model be $\mathbf{X}$, then the inverse of $\mathbf{X}'\mathbf{X}$ is

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.047 & -0.013 & 0 & -0.023 \\ -0.013 & 0.011 & 0 & 0 \\ 0 & 0 & 0.0000051 & 0 \\ -0.023 & 0 & 0 & 0.04 \end{bmatrix}.$$

Selected SAS output is also provided below.

```
                        The GLM Procedure

Dependent Variable: post
***Table One***
                                   Sum of
Source                      DF        Squares    Mean Square   F Value     Pr > F

Model                       ???        644.58        ???         ???         <.0001
Error                       ???        67.24         ???
Corrected Total              99         ???



***Table Two***
Standard Parameter          Estimate           Error        t Value    Pr > |t|
Intercept                   0.75                ???          ???        <0.0001
pre                         1.15                ???          ???        <0.0001
age                         0.052               0.0019       ???        <0.0001
gender                      ???                 ???          6.43       <0.0001
```

Based on this output, answer the following questions:

- *(10 points)* Fill in the cells with ??? in Table One. What are the degrees of freedom associated with the F test?

```
                        The GLM Procedure

Dependent Variable: post
***Table One***
                                   Sum of
Source                      DF        Squares    Mean Square   F Value     Pr > F

Model                       (3)        644.58    (214.86)     (306.9)     <.0001
Error                       (96)        67.24     (0.7)
Corrected Total              99        (711.82)
```

- *(3 points)* Estimate σ^2.
  Solution: $\hat{\sigma}^2 = 0.7$.

- *(5 points)* Report a F-test of the hypothesis that the prior physical activity levels are unrelated to the post-construction physical activity, after adjusting effects of age and gender. Give the nested models implicitly being compared when one conducts this F-test.

Compare full model $post = \beta_0 + \beta_1 pre + \beta_2 age + \beta_3 gender + error$ with $post = \beta_0 + \beta_2 age + \beta_3 gender + error$ or testing $H_0 : \beta_1 = 0$.
t-test $= \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = 1.15/\sqrt{\hat{\sigma}^2 * 0.011} = 13.1 \sim t(96) \approx N(0,1)$ thus significant at $\alpha = 0.05$. Thus F-test $= (t - test)^2 = 13.1^2 = 171.6 \sim F_{1,96}$.

- *(7 points)* Fill in the cells with ??? in Table Two.

Table Two

Standard Parameter	Estimate	Error	t Value	Pr > |t|
Intercept	0.75	(0.181)	(4.14)	<0.0001
pre	1.15	(0.088)	(13.1)	<0.0001
age	0.052	0.0019	(27.4)	<0.0001
gender	(1.07)	(0.167)	6.43	<0.0001

- *(4 points)* Test $H_0 : \beta_1 = 1$.
  Solution: t-test $= \frac{\hat{\beta}_1 - 1}{se(\hat{\beta}_1)} = 0.15/\sqrt{\hat{\sigma}^2 * 0.011} = 1.71 \sim t_{96} \approx N(0,1)$. Since $1.71 < 1.96$, the test is not significant at $\alpha = 0.05$.
- *(8 points)* What is the interpretation of the intercept in the aforementioned regression model? To make β_0 more interpretable, the investigator decides to rescale the variable age by its mean which is 40, and refit the following regression model:
  $post = \beta_0 + \beta_1 pre + \beta_2 newage + \beta_3 gender + error$
  where $newage = age - 40$. Fill in the cells with ??? in Table Three.

***Table Three ***

Standard Parameter	Estimate	Error	t Value	Pr > |t|
Intercept	2.83	0.196	14.4	<0.0001
pre	1.15	0.088	13.1	<0.0001
newage	0.052	0.0019	27.4	<0.000
gender	1.07	0.167	6.43	<0.0001

Solution: The intercept is the expected post construction physical activity per day for a male resident with age 0 and average physical activity per day of 0 hours before the construction.

(3 points) Explain the assumption of homogeneity in the context of this experiment. Is it possible to assess the validity of this assumption from the summary statistics given? If so, how?

Solution: Homogeneity means that variability of the random error is constant across all subjects. No way to assess this assumption without the residuals, or any way to compute them.

1. (25pts) A new drug "B" has been developed to reduce cholesterol level. It was claimed that the new drug is more effective than the old one named "A". In a large scale study, each of these two drugs is tested on 500 patients at 5 doses, with 100 patients per dose, and thus the total sample size is 1000.

   (a) (3pts) First consider the dose variable as a factor with 5 levels, and employ an additive model:

   $$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}.$$

   Using reference cell coding, where $i = 1$, and α_1 models the effect of drug A (drug B is reference); $j = 1, 2, 3, 4$, such that β_j models the effect for dose j (dose 5 is reference); k=1,2, ..., 100, which are patient indices within one cell, and e_{ijk} indicates residual error. If we write this ANOVA model as a regression model: $y = \mathbf{X}b + e$, what is the dimension of y, $\mathbf{X}$, b and e, and for an ANOVA model, what kind of distribution e should follow?

   y: 1000×1

   X: 1000×6

   b: 6×1

   e: 1000×1

   $e \sim N(0, \sigma^2 I)$

   (b) (4pts) For the model specified in part (a), write the cell mean for each combination of drug and dose in terms of μ, α_i and β_j.

Drug	*Dose*	*Mean*
A	1	$\mu+\alpha_1+\beta_1$
A	2	$\mu+\alpha_1+\beta_2$
A	3	$\mu+\alpha_1+\beta_3$
A	4	$\mu+\alpha_1+\beta_4$
A	5	$\mu+\alpha_1$
B	1	$\mu+\beta_1$
B	2	$\mu+\beta_2$
B	3	$\mu+\beta_3$
B	4	$\mu+\beta_4$
B	5	μ

(c) (3pts) For the model specified in part (a), fill the following ANOVA table.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	146250	29250	70.14	<.0001
Error	994	414498	417		
Corrected Total	999	560748			

(d) (3pts) If we model the interaction between dose and drug, the model can be written as

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

where γ_{ij} indicates interaction effects. If we write this ANOVA model as a regression model: $y = \mathbf{X}b + e$, what is the dimension of y, $\mathbf{X}$, b and e

y: 1000×1

X: 1000×10

b: 10×1

e: 1000×1

(e) (4pts) Write the cell mean for each combination of drug and dose in terms of μ, α_i, β_j and γ_{ij}. Explain the meaning of interaction effect γ_{11} by comparing the table in question (b) and the table in this question.

Drug	*Dose*	*Mean*
A	1	$\mu+\alpha_1+\beta_1+\gamma_{11}$
A	2	$\mu+\alpha_1+\beta_2+\gamma_{12}$
A	3	$\mu+\alpha_1+\beta_3+\gamma_{13}$
A	4	$\mu+\alpha_1+\beta_4+\gamma_{14}$
A	5	$\mu+\alpha_1$
B	1	$\mu+\beta_1$
B	2	$\mu+\beta_2$
B	3	$\mu+\beta_3$
B	4	$\mu+\beta_4$
B	5	μ

$$\gamma_{11} = \left(E[y|A,1] - E[y|A,5]\right) - \left(E[y|B,1] - E[y|B,5]\right)$$

γ_{11} is the difference between the difference of dose 1 and dose 5 given drug A and B, respectively.

(f) (4pts) Let μ_A and μ_B be the overall mean values of cholesterol level for drug A and B, respectively. Write down μ_A and μ_B in terms of α_i, β_j and γ_{ij}. If we want to test $H_0 : \mu_A = \mu_B$, write down H_0 in terms of α_i, β_j and γ_{ij}.

$$\mu_A = \frac{(\mu+\alpha_1+\beta_1+\gamma_{11})+(\mu+\alpha_1+\beta_2+\gamma_{12})+(\mu+\alpha_1+\beta_3+\gamma_{13})+(\mu+\alpha_1+\beta_4+\gamma_{14})+(\mu+\alpha_1)}{5}$$

$$\mu_B = \frac{(\mu+\beta_1)+(\mu+\beta_2)+(\mu+\beta_3)+(\mu+\beta_4)+\mu}{5}$$

$$H_0: \mu_A = \mu_B \iff H_0: \alpha_1 + \frac{\gamma_{11}+\gamma_{12}+\gamma_{13}+\gamma_{14}}{5} = 0$$

(g) (3pts) Give an example that $\mu_A = \mu_B$, but the effect of drug A and B are not the same for all the doses.

Suppose $\gamma_{11} = \gamma_{12} = \gamma_{13} = \gamma_{14}/2 = -\alpha_1 \neq 0$

then $E[Y|A,4] = \mu + \alpha_1 + \beta_4 + \gamma_{14}$

$= \mu + \alpha_1 + \beta_4 - 2\alpha_1$

$= \mu + \beta_4 - \alpha_1 \neq E[Y|B,4]$

but clearly $\mu_A = \mu_B$

2. (15pts) Following question 1, we consider to include interval type of variables.

(a) (4pts) Now if we model dose as a interval variable, with doses equals to 1, 2, 3, 4, and 5, and fit a model of cholesterol level with additive effect of dose and drug, but no interaction, fill the following ANOVA table

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	145560	72780	174.95	<.0001
Error	997	414752	416		
Corrected Total	999	560312			

(b) (1pts) Is the model in 2(a) an ANOVA model, an ANCOVA model, or a full model in each cell?

ANCOVA model

(c) (4pts) Compare the model using dose as a categorical variable and the model using dose as a interval variable by F-test. Please write down H_0, calculate F-Statistic, and give the degree of freedom of the corresponding F-distribution when H_0 is true.

$H_0: \beta_1 = 4\beta_4$ & $\beta_2 = 3\beta_4$ & $\beta_3 = 2\beta_4$

$\Leftrightarrow H_0: \beta_1 - 4\beta_4 = 0$ & $\beta_2 - 3\beta_4 = 0$ & $\beta_3 - 2\beta_4 = 0$

$$F\text{-test} = \frac{[SSE(R) - SSE(F)]/3}{SSE[F]/dfE}$$

$$= \frac{(414752 - 414498)/3}{414498/994} = \frac{84.67}{417} = 0.2 \sim F_{3,994}$$

Now we introduce another interval variable "age", and obtained the following output.

Dependent Variable: LDL LDL cholesterol, mg/dL

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	177854.2865	35570.8573	92.38	<.0001
Error	994	382744.6685	385.0550		
Corrected Total	999	560598.9550			

R-Square	Coeff Var	Root MSE	LDL Mean
0.317258	15.32973	19.62282	128.0050

Source	DF	Type I SS	Mean Square	F Value	Pr > F
drug	1	123876.9000	123876.9000	321.71	<.0001
dose	1	21681.7710	21681.7710	56.31	<.0001
age	1	26676.8663	26676.8663	69.28	<.0001
drug*dose	1	2526.5193	2526.5193	6.56	0.0106
drug*age	1	3092.2299	3092.2299	8.03	0.0047

Source	DF	Type III SS	Mean Square	F Value	Pr > F
drug	1	188.590646	188.590646	0.49	0.4842
dose	1	4596.699264	4596.699264	11.94	0.0006
age	1	6103.211364	6103.211364	15.85	<.0001
drug*dose	1	2443.995325	2443.995325	6.35	0.0119
drug*age	1	3092.229910	3092.229910	8.03	0.0047

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	98.82355769	3.53245943	27.98	<.0001
drug	3.55006409	5.07268042	0.70	0.4842
dose	2.14467577	0.62072607	3.46	0.0006
age	0.29584942	0.07431096	3.98	<.0001
drug*dose	2.21124424	0.87770366	2.52	0.0119
drug*age	0.30090703	0.10618369	2.83	0.0047

(d) (2pts) Write down the fitted model based on the above output. Is dose treated as categorical or interval variable? Is this model an ANOVA model, an ANCOVA model, or a full model in each cell?

$$\hat{y} = 98.82 + 3.55\, I\{drug=A\} + 2.145 \cdot dose + 0.296 \cdot age + 2.211\, I\{drug=A\} \cdot dose + 0.3\, I\{drug=A\} \cdot age$$

In this model, dose is treated as an interval variable. This model is a full model in each cell.

(e) (2pts) Write down the fitted model when drug B is used (the reference level for variable drug), using cholesterol level as response, and using ~~drug~~ age and dose as covariates.

$$\hat{y} = 98.82 + 2.145\, dose + 0.296\, age$$

(f) (2pts) Write down the fitted model when drug A is used, using cholesterol level as response, and using ~~drug~~ age and dose as covariates

$$\hat{y} = 102.37 + 4.356\, dose + 0.596\, age$$

(b) (5pts) The result in the previous logistic regression suggest weight is not important, we tried to fit the following smaller model.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	541.990	287.592
SC	545.981	303.557
-2 Log L	539.990	279.592

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	260.3980	3	<.0001
Score	200.4918	3	<.0001
Wald	80.9970	3	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
strain	1	1.6216	0.2029
activity	1	37.1344	<.0001
activity*strain	1	8.3173	0.0039

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-7.2628	1.1901	37.2428	<.0001
strain	B6	1	1.5155	1.1901	1.6216	0.2029
activity		1	1.7819	0.2924	37.1344	<.0001
activity*strain	B6	1	-0.8433	0.2924	8.3173	0.0039

Compared these two models in part (a) and (b) by a likelihood ratio test. Write down H_0, test-statistic, degree of freedom and the distribution of the test statistic when H_0 is true.

H_0: $\beta_{weight} = \beta_{weight*strain} = 0$

LRT [redacted] $-2LR(\text{Reduced}) + 2LR(\text{full})$

$= 279.59 - 279.38$

$= 0.21 \overset{H_0}{\sim} \chi^2_2$

3. (20 pts) In a mouse study, we are interested in tumor occurrences of 400 mice from two strains: 200 mice from B6 and 200 mice from Cast. Mice from one strain all share the same genetic background. This is a regression problem with one response, tumor occurrence, and three predictors: mouse strain (a binary variable), body weight (a continuous/interval variable), and activity index (an continuous/interval variable).

   (a) (5pts) In a simplified situation, we record 1 if a mouse has at least one tumor and 0 otherwise. Then tumor occurrence is a binary variable, and the results of a logistic regression is shown below:

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	541.990	291.381
SC	545.981	315.330
-2 Log L	539.990	279.381

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	260.6084	5	<.0001
Score	200.6430	5	<.0001
Wald	80.9250	5	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-7.2166	1.7079	17.8538	<.0001
strain	B6	1	1.9460	1.7079	1.2983	0.2545
weight		1	-0.00269	0.0606	0.0020	0.9646
activity		1	1.7828	0.2931	36.9886	<.0001
weight*strain	B6	1	-0.0217	0.0606	0.1281	0.7205
activity*strain	B6	1	-0.8427	0.2931	8.2642	0.0040

Please write down the fitted model in the form of $E(y_i) = f(\hat{\beta})$ based on the above SAS output, where $\hat{\beta}$ are the regression coefficient estimates. What is $Var(y_i)$?

$$E(y_i) = f(\hat{\beta}) = \hat{p} = \frac{\exp\{g(\hat{\beta})\}}{1+\exp\{g(\hat{\beta})\}}$$

Where $g(\hat{\beta}) = -7.22 + 1.95\, I\{\text{strain} = B6\} - 0.0027\,\text{weight} + 1.78\,\text{activity} - 0.0217\, I\{\text{strain} = B6\}\cdot\text{weight} - 0.8427\, I\{\text{strain} = B6\}\cdot\text{activity}$

$$Var(y_i) = \hat{p}(1-\hat{p}) = \frac{\exp\{g(\hat{\beta})\}}{(1+\exp\{g(\hat{\beta})\})^2}$$

In a follow-up study, we took 20 mice with tumor (10 from strain B6 and 10 from Cast) and 20 mice without tumor (10 B6 + 10 Cast), and measure the expression of a gene that is important in tumor progression at three tissues of each mouse: left forebrain, left hindbrain, and right whole brain. We have altogether (20+20)*3 = 120 measurements of gene expression.

(c) (2pts) Please describe the structure of the 120*120 covariance matrix of these 120 observations. How many elements of this matrix are expected to be 0?

block diagonal

$120 \times 120 - 3 \times 3 \times 40$

$= 14040$ elements are expected to be 0

(d) (2pts) Here are the results of one mixed effect model, what kind of covariance structure are assumed for three expression measurements per mouse?

Estimated R Matrix for mouseID 1

Row	Col1	Col2	Col3
1	2.1015	0.6881	0.6881
2	0.6881	2.1015	0.6881
3	0.6881	0.6881	2.1015

Fit Statistics

-2 Res Log Likelihood	417.4
AIC (smaller is better)	421.4
AICC (smaller is better)	421.5
BIC (smaller is better)	424.8

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	11.11	0.0009

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
tumor	1	37	4.43	0.0421
strain	1	37	22.02	<.0001

why type I SS and type III SS in the following output are the same.

Dependent Variable: expression

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	91.9864678	45.9932339	22.26	<.0001
Error	117	241.7472351	2.0662157		
Corrected Total	119	333.7337030			

R-Square	Coeff Var	Root MSE	expression Mean
0.275628	158.7455	1.437434	0.905524

Source	DF	Type I SS	Mean Square	F Value	Pr > F
tumor	1	15.41973044	15.41973044	7.46	0.0073
strain	1	76.56673739	76.56673739	37.06	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
tumor	1	15.41973044	15.41973044	7.46	0.0073
strain	1	76.56673739	76.56673739	37.06	<.0001

① the model violates the independence assumption!

② the p value gets small due to the independence assumption violation [redacted]

③ since the ~~expo~~ data is balanced & design matrix corresponding to tumor & strain are orthogonal.

In a follow-up study, we took 20 mice with tumor (10 from strain B6 and 10 from Cast) and 20 mice without tumor (10 B6 + 10 Cast), and measure the expression of a gene that is important in tumor progression at three tissues of each mouse: left forebrain, left hindbrain, and right whole brain. We have altogether (20+20)*3 = 120 measurements of gene expression.

(c) (2pts) Please describe the structure of the 120*120 covariance matrix of these 120 observations. How many elements of this matrix are expected to be 0?

(d) (2pts) Here are the results of one mixed effect model, what kind of covariance structure are assumed for three expression measurements per mouse?

Compound Symmetry

Estimated R Matrix for mouseID 1

Row	Col1	Col2	Col3
1	2.1015	0.6881	0.6881
2	0.6881	2.1015	0.6881
3	0.6881	0.6881	2.1015

Fit Statistics

-2 Res Log Likelihood	417.4
AIC (smaller is better)	421.4
AICC (smaller is better)	421.5
BIC (smaller is better)	424.8

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	11.11	0.0009

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
tumor	1	37	4.43	0.0421
strain	1	37	22.02	<.0001

unstructured

(e) (3pts) Here are the results of the other mixed effect model, what kind of covariance structure are assumed for the three expression measurements per mouse in this model? Compare this model with previous one by a Likelihood Ratio test, write down test statistic, degree of freedom and the distribution of the test statistic when Null hypothesis is correct.

The Mixed Procedure

Estimated R Matrix for mouseID 1

Row	Col1	Col2	Col3
1	2.4998	1.3469	0.1251
2	1.3469	1.9588	0.5887
3	0.1251	0.5887	1.8423

Fit Statistics

-2 Res Log Likelihood	404.3
AIC (smaller is better)	416.3
AICC (smaller is better)	417.1
BIC (smaller is better)	426.5

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
5	24.21	0.0002

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
tumor	1	37	4.22	0.0471
strain	1	37	23.26	<.0001

$LRT = 417.4 - 404.3$

$= 13.1 \sim \chi^2_4$

(f) (3pts) Someone ignored the fact that these mouse are not independent and did a fixed effect linear regression. Compared the following results with the results from question (e), explain (i) which assumption of general linear regression is violated, (ii) why we see smaller p-values in the fixed effect linear model? (iii) Give a reasonable guess

1. *(40 points total)* A group of subjects was recruited to a nutritional study in a medical center at UNC. The data consist of their BMI (y = BMI), daily exercise time (x1 = exercise (in hours)) and daily vegetable intake (x2 = vegetable (in servings)). One of the objectives in this study is to estimate how the exercise and vegetable consumption affect BMI. To address the question, we consider the following model:

$$y = \beta_0 + \beta_1 x1 + \beta_2 x2 + \epsilon$$

. Let $\mathbf{X}$ be the associated design matrix of the above model. The data is summarized below:

$$\mathbf{X'X} = \begin{pmatrix} 200.0000 & 588.7676 & 1033.797 \\ 588.7676 & 2321.7635 & 2951.400 \\ 1033.7973 & 2951.3999 & 7138.232 \end{pmatrix}, \mathbf{X'y} = \begin{pmatrix} 4647.273 \\ 13561.768 \\ 23709.514 \end{pmatrix},$$

and $(\mathbf{X'X})^{-1} = \begin{pmatrix} 0.037523205 & -0.005495959 & -0.003161934 \\ -0.005495959 & 0.001712864 & 0.00008774738 \\ -0.003161934 & 0.00008774738 & 0.0005617387 \end{pmatrix}$.

- (8 points) A partial ANOVA table is given below. Complete the table.

The GLM Procedure
Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	? 2	85.543209	? 42.77	? 7.65	-
Error	? 197	1101.480037	? 5.59		
Corrected Total	199	1187.023246			

- (8 points) Compute the least square estimates of the model parameters and their standard errors. Conduct the tests for the significance of each parameter (i.e., $H0 : \beta_1 = 0$, and $H0 : \beta_2 = 0$).

$\hat{\beta} = (X'X)^{-1}X'y = \begin{pmatrix} 24.878 \\ -0.231 \\ -0.1858 \end{pmatrix}$ $\quad \widehat{Cov(\hat{\beta})} = \hat{\sigma}^2(X'X)^{-1}$

$\hat{\sigma}^2 = MSE = 5.59$

For $H_0: \beta_1 = 0$

$T = \frac{\hat{\beta}_1}{se(\hat{\beta})} = \frac{-0.231}{\sqrt{0.001712 \times 5.59}} = -2.36 \overset{H_0}{\sim} t_{197}$ & both $|t| > 1.96$ so reject H_0 at 0.05.

For $H_0: \beta_2 = 0$

$T = \frac{-0.1858}{\sqrt{0.0005617 \times 5.59}} = -3.32 \overset{H_0}{\sim} t_{197}$

- (8 points) Compute the 95% confidence interval for the BMI of individuals who on average exercise 2 hours and eat 6 servings of vegetables daily.

$\beta^* = \beta_0 + 2\beta_1 + 6\beta_2$

$\Rightarrow \hat{\beta}^* = \hat{\beta}_0 + 2\hat{\beta}_1 + 6\hat{\beta}_2 = 23.3$

$\widehat{Var}(\hat{\beta}^*) = (1 \quad 2 \quad 6)\, Var(\hat{\beta}) \begin{pmatrix} 1 \\ 2 \\ 6 \end{pmatrix} = 0.03789$

$\therefore$ CI of β^* is $\hat{\beta}^* \pm 1.96\, se(\hat{\beta}^*)$

$= (22.92,\ 23.68)$

- (8 points) Test $H_0 : \beta_1 = 3\beta_2$.

Let $\theta = \beta_1 - 3\beta_2$ then $H_0 : \theta = 0$

$$t = \frac{\hat{\theta}}{se(\hat{\theta})} = \frac{(0\ 1\ -3)\hat{\beta}}{\sqrt{\hat{\sigma}^2 (0\ 1\ -3)(X'X)^{-1} \begin{pmatrix} 0 \\ 1 \\ -3 \end{pmatrix}}}$$

$$= \frac{0.327}{0.1868} = 1.75 \sim t_{197}$$

cannot reject H_0 since $|t| < 1.96$

- (8 points) Next we center the exercise and vegetable consumptions at their means, which are 1 hour and 5 servings respectively and refit the data with the new transformed variables. Fill in the cells with ? in the following table.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	?23.7174	? 0.2141	? 93.37	-
newx1	?-0.231	? 0.0979	? -2.36	-
newx2	?-0.186	? 0.0596	? -3.13	-

$newX1 = X1 - 1$

$newX2 = X2 - 5$

so if original model is

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$

then new model should be

$y = \beta_0^* + \beta_1^* newx_1 + \beta_2^* newx_2 + e$

with $\beta_0^* = \beta_0 + \beta_1 + 5\beta_2$

$\beta_1^* = \beta_1$

$\beta_2^* = \beta_2$

2. *(40 points total)* This study investigates how the four dose levels of Vitamin C (1, 2, 3 and 4 mg) and two delivery methods (orange juice or ascorbic acid) affect the length of odontoblasts (teeth) in 800 guinea pigs. The study is balanced, so for each dose and delivery method combination, 100 pigs are assigned.

   - (14 points) first consider the dose variable as categorial and employ an additive model using reference cell coding (where ascorbic acid and dosage 1mg are used as references respectively):

$$y_i = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \beta x_{4i} + e_i$$

   where α_is refer the dosage effects and β refers the effect of delivery method. Describe dummy variables x_{1i}, x_{2i}, x_{3i} and x_{4i}, based on which write down the cell mean of each group in terms of μ, α_is and β in the following table.

$x_{1i} = \begin{cases} 1 & \text{if the dose level is 2} \\ 0 & \text{if the dose level} \neq 2 \end{cases}$

$x_{2i} = \begin{cases} 1 & \text{if dose level} = 3 \\ 0 & \text{otherwise} \end{cases}$

$x_{3i} = \begin{cases} 1 & \text{if dose level} = 3 \\ 0 & \text{otherwise} \end{cases}$

$x_{4i} = \begin{cases} 1 & \text{if delivery method is orange juice} \\ 0 & \text{otherwise} \end{cases}$

Delivery method	Dose	Mean
Orange juice	1	$\mu+\beta$
Orange juice	2	$\mu+\alpha_1+\beta$
Orange juice	3	$\mu+\alpha_2+\beta$
Orange juice	4	$\mu+\alpha_3+\beta$
Ascorbic acid	1	μ
Ascorbic acid	2	$\mu+\alpha_1$
Ascorbic acid	3	$\mu+\alpha_2$
Ascorbic acid	4	$\mu+\alpha_3$

   - (7 points) If we add the interaction terms between the delivery methods and dosage into the above model and express the new model in matrix notation $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{e}$, what are the dimensions of $\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}$ and $\mathbf{e}$?

$y_{800\times 1}$; $X: 800\times 8$; $\theta = 8\times 1$

$e: 800\times 1$

- (12 points) Let μ_{orange} and $\mu_{ascorbic}$ be the overall means of the two delivery methods. Write down μ_{orange} and $\mu_{ascorbic}$ for the models with and without interaction terms. Derive the two **C** matrices for testing $H_0 : \mu_{orange} = 2\mu_{ascorbic}$ under the two models.

Without interaction: $\mu_{orange} = \mu + \frac{\alpha_1+\alpha_2+\alpha_3}{4} + \beta$

$\mu_{ascorbic} = \mu + \frac{\alpha_1+\alpha_2+\alpha_3}{4}$

So $H_0: \mu_{orange} = 2\mu_{ascorbic} \iff \mu + \frac{\alpha_1+\alpha_2+\alpha_3}{4} - \beta = 0$

i.e. $C\theta = (1\ \frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}\ -1)$ for $\theta = (\mu, \alpha_1, \alpha_2, \alpha_3, \beta)^T$

With interaction: $\mu_{orang} = \mu + \frac{\alpha_1+\alpha_2+\alpha_3}{4} + \beta + \frac{\gamma_{11}+\gamma_{12}+\gamma_{13}}{4}$

$\mu_{ascorbic} = \mu + \frac{\alpha_1+\alpha_2+\alpha_3}{4}$

$H_0: \mu_{orange} = 2\mu_{ascorbic} \iff \mu + \frac{\alpha_1+\alpha_2+\alpha_3}{4} - \beta - \frac{\gamma_{11}+\gamma_{12}+\gamma_{13}}{4} = 0$

i.e. $C = (1, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, -1, -\frac{1}{4}, -\frac{1}{4}, -\frac{1}{4})^T$ for $\theta = (\mu, \alpha_1, \alpha_2, \alpha_3, \beta, \gamma_{11}, \gamma_{12}, \gamma_{13})$

- (7 points) Next, treat the Vitamin C dosage as a continuous variable and fit a model with additive effects of the delivery method and vitamin C level, with no interaction. Is this model nested within Model

$$y_i = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \beta x_{4i} + e_i?$$

If yes, write down H_0 for comparing the two models and derive **C** matrix. What are the degrees of freedom of the corresponding F test under H_0?

yes When letting $\alpha_2 = 2\alpha_1$ and $\alpha_3 = 3\alpha_1$ we basically assume that the dosage level as a continuous variable.

So the C matrix assuming $\theta = (\mu, \alpha_1, \alpha_2, \alpha_3, \beta)^T$

$$C = \begin{bmatrix} 0 & -2 & 1 & 0 & 0 \\ 0 & -3 & 0 & 1 & 0 \end{bmatrix}$$

the degrees of freedom of F are 2, 795.

3. *(20 points total)* Table below lists a data set derived from a study on the relationship between incubation temperature for hatching turtle eggs and gender of baby turtles.

Temperature	Male	Female	Total
low	10	40	50
medium	28	22	50
high	34	16	50

To study how the incubation temperature affects the sex of baby turtles, we fit the logistic regression model

$$logit(p) = \mu + \beta_1 I(temperature = low) + \beta_2 I(temperature = medium)$$

where p is the probability of hatching a male turtle, and get the following output:

```
                     Estimate   Std. Error  z value Pr(>|z|)
intercept             0.7538     0.3032     2.486   0.0129
I(temperature=low)   -2.1401     0.4657    -4.595   4.33e-06
I(temperature=medium) -0.5126    0.4160    -1.232   0.2179
```

- (10 points) Estimate the probability that a male turtle hatches from an egg incubated at medium temperature.

$$\log \frac{\hat{p}}{1-\hat{p}} = 0.7538 - 0.5126$$

$$\Rightarrow \hat{p} = 0.56$$

- (5 points) What is the estimate of the odds ratio of low vs high temperatures and construct a 95% confidence interval for this odds ratio.

$$\log(OR) = \log\left(\frac{\frac{\hat{p}_{low}}{1-\hat{p}_{low}}}{\frac{\hat{p}_{high}}{1-\hat{p}_{high}}}\right) = (\hat{\mu} + \hat{\beta}_1) - (\hat{\mu})$$

$$= \hat{\beta}_1 = -2.1401$$

So CI of OR is

$$[\exp(\hat{\beta}_1 - 1.96 \times 0.4657), \exp(\hat{\beta}_1 + 1.96 \times 0.4657)]$$

$$= [0.0472, 0.293]$$

- (5 points) What is the estimate of the odds ratio of low vs medium temperatures. Do you have enough information to construct a 95% confidence interval for this odds ratio? If yes, construct the CI. If not, explain why.

point estimate

$$\widehat{OR} = \exp\{\hat{\mu} + \hat{\beta}_1 - \hat{\mu} - \hat{\beta}_2\}$$

$$= \exp\{\hat{\beta}_1 - \hat{\beta}_2\} = \exp\{-2.1401 + 0.5726\}$$

$$= 0.196$$

to get confidence interval, we need $cov(\hat{\beta}_1, \hat{\beta}_2)$ which is not available to us. So cannot get the CI.

1. (28pts) Consider the model $\mathbf{y}_{8\times1} = \mathbf{X}_{8\times3}\boldsymbol{\beta}_{3\times1} + \boldsymbol{\epsilon}_{8\times1}$, where $\mathbf{y}$ is blood pressure of 8 individuals, $\mathbf{X}$ includes intercept (1st column of $\mathbf{X}$) and two covariates: age (2nd column of $\mathbf{X}$) and body weight (lbs) (3rd column of $\mathbf{X}$). More specifically,

BP intercept age bwt

$$\mathbf{y} = \begin{bmatrix} 137 \\ 126 \\ 114 \\ 95 \\ 111 \\ 112 \\ 107 \\ 121 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 26 & 134 \\ 1 & 27 & 138 \\ 1 & 23 & 118 \\ 1 & 24 & 124 \\ 1 & 22 & 123 \\ 1 & 30 & 135 \\ 1 & 20 & 128 \\ 1 & 25 & 131 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \text{and } \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_8 \end{bmatrix} \sim N(0, \sigma^2\mathbf{I})$$

You should NOT run any software to answer the following questions. However, some computation by calculator maybe needed given the following potential helpful facts.

- The corrected total sum of squares of $\mathbf{y}$ is 1476.
- $(\mathbf{X}^T\mathbf{X})^{-1} =$

	intercept	age	weight
intercept	57.406	0.435	-0.528
age	0.435	0.028	-0.009
weight	-0.528	-0.009	0.006

- $\hat{\sigma}^2 = 145.37$.

(a) (5pts) Is each of the following statement correct or not? If it is not correct, please explain why it is wrong and try to correct it.

i. β are statistics.

Incorrect, β's are parameters that we can't observe, we use $\hat{\beta}$ to estimate them.

ii. ϵ are parameters.

Incorrect, ϵ's are random errors.

iii. $\mathbf{y}$ is a random variable following multivariate normal distribution with mean value $\mathbf{0}_{8\times1}$ and variance $\sigma^2\mathbf{I}_{8\times8}$.

Incorrect, y is a random variable following multivariate normal distribution but the mean value $E(y) = X\beta$, not $E(\epsilon)$, covariance $= \sigma^2 I_{8\times8}$

iv. $\hat{\sigma}^2$ is a random variable.

Correct. $\hat{\sigma}^2$ is the estimator of σ^2 and is a random variable.

v. ϵ_1 is independent with ϵ_2.

Correct. random errors are assumed to be independent of each other.

(b) (3pts) Fill in the following t-table and please show your work on calculating the Standard Errors.

Parameter	Estimate	Standard Error	t value	Pr(>|t|)
(Intercept)	-22.0801	91.3516	-0.241	0.823
age	-0.1105	2.0175	-0.0548	0.959
weight	1.0877	0.9339	1.1647	0.299

$Cov(\hat{\beta}) = \sigma^2 (X'X)^{-1}$

$= \hat{\sigma}^2 (X'X)^{-1}$

$= 145.37 \begin{bmatrix} 57.406 & 0.435 & -0.528 \\ 0.435 & 0.028 & -0.009 \\ -0.528 & -0.009 & 0.006 \end{bmatrix}$

$= \begin{bmatrix} 8345.11 & 63.23575 & -76.7554 \\ 63.23575 & 4.07036 & -1.30833 \\ -76.7554 & -1.30833 & 0.87222 \end{bmatrix}$

$Var(\hat{\beta}_0) = 8345.11 \quad Se(\hat{\beta}_0) = \sqrt{8345.11} = 91.3516$

$Var(\hat{\beta}_1) = 4.07036 \quad Se(\hat{\beta}_1) = \sqrt{4.07036} = 2.0175$

$Var(\hat{\beta}_2) = 0.87222 \quad Se(\hat{\beta}_2) = \sqrt{0.87222} = 0.9339$

$t_{\hat{\beta}_0} = \frac{-22.0801 - 0}{91.3516} = -0.241$

$t_{\hat{\beta}_1} = \frac{-0.1105 - 0}{2.0175} = -0.0548$

$t_{\hat{\beta}_2} = \frac{1.0877 - 0}{0.9339} = 1.164$

(c) (5pts) Test $\beta_0 = \beta_1 = \beta_2$ using GLH approach. Write out the contrast matrix **C**, calculate test statistic and specify its null distribution and the corresponding degree of freedom. Though you do not need to calculate the p-value.

$\beta_0 = \beta_1 \Rightarrow \beta_0 - \beta_1 = 0$

$\beta_1 = \beta_2 \Rightarrow \beta_1 - \beta_2 = 0 \Rightarrow C = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$

$H_0: \theta = \begin{bmatrix} \beta_0 - \beta_1 \\ \beta_1 - \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, corresponding contrast matrix $C = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$

Because X is full rank, so θ is estimable.

Also since C is full rank, so θ is testable.

$M_{2\times 2} = C(X'X)^{-1}C' = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 57.406 & 0.435 & -0.528 \\ 0.435 & 0.028 & -0.009 \\ -0.528 & -0.009 & 0.006 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}$

$= \begin{bmatrix} 58.504 & 0.926 \\ 0.926 & 0.052 \end{bmatrix}$

$M^{-1} = \begin{bmatrix} 0.025 & -0.444 \\ -0.444 & 27.144 \end{bmatrix}$

$\hat{\theta} = C\hat{\beta} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} -22.0801 \\ -0.1105 \\ 1.0877 \end{bmatrix} = \begin{bmatrix} -21.9696 \\ -1.1982 \end{bmatrix}$

degree of freedom: 2, 5

$F\text{-obs} = \frac{(\hat{\theta} - \theta_0)' M^{-1} (\hat{\theta} - \theta_0)/a}{\hat{\sigma}^2} = \frac{[-21.9696 \ -1.1982] \begin{bmatrix} 0.025 & -0.444 \\ -0.444 & 27.144 \end{bmatrix} \begin{bmatrix} -21.9696 \\ -1.1982 \end{bmatrix} / 2}{145.37} = 0.095$

(d) (5pts) Test $\beta_1 = \beta_2 = 0$ using GLH approach. Write out the contrast matrix C, calculate test statistic and specify its null distribution and the degree of freedom. Though you do not need to calculate the p-value.

H_0: $\theta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, corresponding contrast matrix $C = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

Because X is full rank, C is full rank, so θ is testable.

$M_{2\times 2} = C(X'X)^{-1}C' = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 57.906 & 0.435 & -0.528 \\ 0.435 & 0.028 & -0.009 \\ -0.528 & -0.009 & 0.006 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0.028 & -0.009 \\ -0.009 & 0.006 \end{bmatrix}$

$M^{-1} = \begin{bmatrix} 68.966 & 103.448 \\ 103.448 & 321.839 \end{bmatrix}$ $\quad \hat{\theta} = C\hat{\beta} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} -22.801 \\ -0.1105 \\ 1.0877 \end{bmatrix} = \begin{bmatrix} -0.1105 \\ 1.0877 \end{bmatrix}$

$F_{obs} = \dfrac{(\hat{\theta}-\theta_0)'M^{-1}(\hat{\theta}-\theta_0)/q}{\hat{\sigma}^2} = \dfrac{\begin{bmatrix} -0.1105 & 1.0877 \end{bmatrix} \begin{bmatrix} 68.966 & 103.448 \\ 103.448 & 321.839 \end{bmatrix} \begin{bmatrix} -0.1105 \\ 1.0877 \end{bmatrix} / 2}{145.37} = \dfrac{356.7891/2}{145.37} = 1.23$

Df: 2, 5

(e) (5pts) Calculate the correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$.

$\text{Correlation} = \dfrac{\text{Cov}(\hat{\beta}_0, \hat{\beta}_1)}{\sqrt{\text{Var}(\hat{\beta}_0)\text{Var}(\hat{\beta}_1)}} = \dfrac{63.23595}{\sqrt{8345.111 \times 4.07036}} = 0.3431$

(f) (5pts) What is the interpretation of β_0, β_1, and β_2, respectively. Is the interpretation of β_0 meaningful, if so, why? If not, how to fix this problem?

β_0 — the expected blood pressure when age and body weight the value zero.

β_1 — the expected increase in blood pressure for one unit increase in age.

β_2 — the expected increase in blood pressure for one unit increase in body weight.

The interpretation of β_0 is not meaningful because of no biological meaning for BP with age = 0, body weight = 0. To fix the problem, we can center the age variable and weight variable by subtracting the average of age and body weight from each observation respectively. In doing so, the intercept β_0 will be the expected blood pressure when age is at the observed average and body weight is at the observed average value.

2. (20pts) Still use the data presented in problem 1. Suppose we are interested in the event of whether blood pressure is larger than 120. Let $\tilde{y}_i = 1$, if $y_i > 120$, and $\tilde{y}_i = 0$ otherwise. Here $i = 1, 2, ..., 8$ is the index of the 8 individuals. Let $p_i = Pr(y_i > 120)$.

   (a) (5pts) Is p_i a parameter or a statistic? Given p_i, what the distribution of $\tilde{y}_i$? Calculate $\tilde{y}_i$'s expectation and variance.

p_i is a parameter

$\tilde{y}_i = \begin{cases} 1, & Pr = p_i \\ 0, & Pr = 1-p_i \end{cases}$ Given p_i, $\tilde{y}_i \sim$ Bernoulli(p_i)

$E(\tilde{y}_i) = p_i$

$Var(\tilde{y}_i) = p_i(1-p_i)$

   (b) (5pts) Calculate the odds ratio of the event $y_i > 120$ vs. the event weight > 132.

For $y_i > 120$, $\frac{p_1}{1-p_1} = \frac{3/8}{5/8} = \frac{3}{5} = 0.60$

For weight > 132, $\frac{p_0}{1-p_0} = \frac{3/8}{5/8} = \frac{3}{5} = 0.60$

$OR = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \frac{0.60}{0.60} = 1$

   (c) (5pts) Now we fit a logistic regression to study the relation $\tilde{y}$ and age and weight. Please use the following regression coefficients estimates,

	Estimate	Std. Error
(Intercept)	-118.9085	135.9180
age	-0.7111	0.9114
weight	1.0373	1.1950

But I count this as correct, by what I meant is

weight	>120	≤120	
>132	2	1	3
≤132	1	4	5
	3	5	

$\text{odd ratio} = \frac{\frac{2/3}{1-2/3}}{(1/5)/(4/5)} = 8$

to estimate the probability that blood pressure is larger than 120 for an individual of age 30 and weight 133.

$$p = \frac{\exp(\beta_0+\beta_1 age+\beta_2 wt)}{1+\exp(\beta_0+\beta_1 age+\beta_2 wt)} = \frac{\exp(-118.9085+(-0.7111)\times 30+1.0313\times 133)}{1+\exp(-118.9085+(-0.7111)\times 30+1.0313\times 133)}$$

$$= 0.093$$

(d) (5pts) Please use the regression coefficient estimates in part (c) to calculate the odds ratio of the event $y_i > 120$ for person B vs. person A. They are of the same age, but B is 10 pounds heavier than A.

$$\log(odds_B) = \hat{\beta}_0 + \hat{\beta}_1 \times age_B + \hat{\beta}_2 \times wt_B$$

$$\log(odds_A) = \hat{\beta}_0 + \hat{\beta}_1 \times age_A + \hat{\beta}_2 \times wt_A, \quad age_B = age_A, \; wt_B = 10 + wt_A$$

$$\log(OR_{B\text{ vs }A}) = \log(odds_B) - \log(odds_A) = \hat{\beta}_2 (wt_B - wt_A)$$

$$= \hat{\beta}_2 \times 10$$

$$OR_{B\text{ vs }A} = e^{10\hat{\beta}_2} = e^{10(1.0313)} = 31984$$

3. (12pts) Now suppose we know the 8 individuals are from two family. The first four are from one family and the next four are from the other family. In order to accommodate the correlations between individuals within one family, we decide to use a random effect model to study the relation between blood pressure versus age and weight.

$Y_{ij} = X_{ij}\beta + b_i + \varepsilon_{ij}$

two families $i=2$

four from a family $j=4$

(a) (4pts) If we use "unstructured" covariance structure, how many parameters of the covariance matrix of the 8 individuals need to be estimated? Write out the covariance matrix using concise notations (you just need to present the form of the matrix, but do not need to calculate the actual values of the matrix elements).

unstructured covariance matrix in one family:

$$\begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{bmatrix}_{4\times 4}$$

$\frac{4\times(4+1)}{2} = 10$ unique elements need to be estimated

For all individuals in the study

$$Cov = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & 0 & 0 & 0 & 0 \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} & 0 & 0 & 0 & 0 \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} & 0 & 0 & 0 & 0 \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ 0 & 0 & 0 & 0 & \sigma_{12} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ 0 & 0 & 0 & 0 & \sigma_{13} & \sigma_{23} & \sigma_{33} & \sigma_{34} \\ 0 & 0 & 0 & 0 & \sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_{44} \end{bmatrix}$$

(b) (4pts) If we used "compound symmetry" covariance structure, how many parameters of the covariance matrix of the 8 individuals need to be estimated? Write out the covariance matrix using concise notations.

Using compound symmetry covariance structure, we need to estimate 2 parameters: σ_b^2 and σ_w^2. For one family:

$$CS = \begin{bmatrix} \sigma_b^2+\sigma_w^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2+\sigma_w^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_w^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_w^2 \end{bmatrix}$$

For all 8 individuals:

$$CS = \begin{bmatrix} \sigma_b^2+\sigma_w^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & 0 & 0 & 0 & 0 \\ \sigma_b^2 & \sigma_b^2+\sigma_w^2 & \sigma_b^2 & \sigma_b^2 & 0 & 0 & 0 & 0 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_w^2 & \sigma_b^2 & 0 & 0 & 0 & 0 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_w^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_b^2+\sigma_w^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ 0 & 0 & 0 & 0 & \sigma_b^2 & \sigma_b^2+\sigma_w^2 & \sigma_b^2 & \sigma_b^2 \\ 0 & 0 & 0 & 0 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_w^2 & \sigma_b^2 \\ 0 & 0 & 0 & 0 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_w^2 \end{bmatrix}_{8\times 8}$$

(c) (2pts) Which covariance structure (unstructured or compound symmetric) should we use for this dataset and why?

Because of the small sample size in this dataset, we should use compound symmetric covariance matrix because it has fewer parameters than unstructured. If assumption for compound symmetry is not valid, we may need to force a compound symmetry structure with appropriate methods.

(d) (2pts) Mixed model parameters can be estimated using either Maximum Likelihood (ML) method or Restricted maximum likelihood (REML) method. In order to compare a model with fixed effects of age and weight vs. the other model with only one fixed effect weight, should we use ML or REML method, and why? (Assume the same covariance structure is used both models.)

We should use ML to compare the two models because the likelihood obtained for models with different fixed effects are not comparable when REML is used to estimate the models. REML maximizes the likelihood of the observed residuals, so different degrees of freedom btw two models, thus they're not comparable.

4. (25pts) We want to compare two drugs (denoted by A and B) for their effects of reducing cholesterol levels (LDL, in the unit of mg/dL). The following table shows the sample size for each combination of drug and dosage.

Drug	Dose	Sample Size (n_{ij})	i (drug index)	j (dose index)
A	1	100	1	1
	2	100	1	2
	3	100	1	3
B	1	100	2	1
	2	100	2	2
	3	100	2	3

(a) (3pts) First consider the dose variable as a categorical variable with 3 levels, and employ an additive model:

drug dose

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk},$$

where i=1, j=1,2, k=1, 2., ..., n_{ij}. We use reference cell coding with drug B and dose 3 as reference. Therefore α_1 models the effect of drug A (drug B is reference), β_j models the effect for dose j (j=1 or 2) (dose 3 is reference); and e_{ijk} (k=1, 2., ..., n_{ij}) indicates residual error. If we write this ANOVA model as a regression model: $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, what is the dimension of $\mathbf{y}$, $\mathbf{X}$, $\mathbf{b}$ and $\mathbf{e}$, and for an ANOVA model, what kind of distribution we usually assume $\mathbf{e}$ should follow?

y = drug A dose 1 dose 2

$y_{600\times1}$, $X_{100\times4}$, $b_{4\times1}$, $e_{600\times1}$.

e follows a Gaussian distribution within cell.

100 × 4 ? — 1

should be 600 × 4

(b) (4pts) For the model specified in part (a), write the cell mean for each combination of drug and dose in terms of μ, α_i and β_j.

Drug	*Dose*	*Mean*
A	1	$\mu+\alpha_1+\beta_1$
A	2	$\mu+\alpha_1+\beta_2$
A	3	$\mu+\alpha_1$
B	1	$\mu+\beta_1$
B	2	$\mu+\beta_2$
B	3	μ

y=drug A dose 1 dose 2

(c) (3pts) For the model specified in part (a), fill the following ANOVA table.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	67284.9	22428.3	57.186	<.0001
Error	596	233751.2	392.2		
Corrected Total	599	301036.1			

(d) (3pts) If we model the interaction between dose and drug, the model can be written as

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

y=drugA dose1 dose2 dose1d1 dose2d2 where γ_{ij} indicates interaction effects. Write the cell mean for each combination of drug and dose in terms of μ, α_i, β_j and γ_{ij}. Explain the meaning of interaction effect γ_{11} by comparing the table in question (b) and the table in this question.

Drug	*Dose*	*Mean*
A	1	$\mu+\alpha_1+\beta_1+\gamma_{11}$
A	2	$\mu+\alpha_1+\beta_2+\gamma_{12}$
A	3	$\mu+\alpha_1$
B	1	$\mu+\beta_1$
B	2	$\mu+\beta_2$
B	3	μ

γ_{11} — the difference in drug effect for dose 1 versus dose 3.

(e) (2pts) Now if we model dose as a interval variable, with doses equals to 1, 2, 3 and fit a model of LDL with main effects of dose and drug, but no interaction, fill the following ANOVA table

y = drug A dose

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	67203.6	33601.8	85.785	<.0001
Error	597	233844.9	391.7		
Corrected Total	599	301048.5			

categorical $y = \mu + \alpha_1 + \beta_1 + \beta_2$

$H_0: \beta_2 = 2\beta_1$

(f) (3pts) Compare the model using dose as a categorical variable (part (c)) and the model using dose as a interval variable (part (e)) by F-test. Please write down H_0, calculate F-Statistic, and give the degree of freedom of the corresponding F-distribution when H_0 is true. Though you do not need to calculate the p-value.

categorical

We can view the model using dose as an interval variable as a model nested in the categorical dose parameterization model.

$y = \mu + \alpha_i$ drug $+ \beta_1$ (dose=1) $+ \beta_2$ (dose=2)

$H_0: \beta_2 = 0$

$$F_{obs} = \frac{\frac{SSE(I) - SSE(C)}{df_E(I) - df_E(C)}}{SSE(C)/df_E(C)} = \frac{\frac{2338449 - 2337512}{597 - 596}}{2337512/596} = 0.2389$$

numerial/interval

$df = 1, 596$

$y = \mu + \alpha_i$ drug $+ \beta_3$ dose

(g) (4pts) Let μ_A and μ_B be the overall mean values of LDL for drug A and B, respectively. Write μ_A and μ_B in terms of α_i, β_j and γ_{ij}. If we want to test $H_0 : \mu_A = \mu_B$, write H_0 in terms of α_i, β_j and γ_{ij}, the contrast matrix, and the degrees of freedom.

dose categorical

$$\mu_A: \frac{(\mu + \alpha_1 + \beta_1 + \gamma_{11}) + (\mu + \alpha_1 + \beta_2 + \gamma_{12}) + (\mu + \alpha_1)}{3}$$

$$= \mu + \alpha_1 + \frac{\beta_1 + \beta_2 + \gamma_{11} + \gamma_{12}}{3}$$

$$\mu_B: \frac{(\mu + \beta_1) + (\mu + \beta_2) + \mu}{3} = \mu + \frac{\beta_1 + \beta_2}{3}$$

$H_0: \mu_A = \mu_B \Rightarrow \mu + \alpha_1 + \frac{\beta_1 + \beta_2 + \gamma_{11} + \gamma_{12}}{3} = \mu + \frac{\beta_1 + \beta_2}{3}$

$\alpha_1 + \frac{\gamma_{11} + \gamma_{12}}{3} = 0$

dose=1	dose=2
$\mu + \alpha_1 + \beta_1$	$\mu + \alpha_1$
	β_2
$\mu + \alpha_1 + \beta_3$	$\mu + \alpha_1 + 2\beta_3$

(categorical) $C = \begin{bmatrix} 0 & 1 & 0 & 0 & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$

interval $df: 1, 594$

(h) (3pts) If the design is unbalanced, with sample size shown in the following table. Test $H_0 : \mu_A = \mu_B$. Write H_0 in terms of α_i, β_j and γ_{ij}, the contrast matrix, and the degrees of freedom. dose categorical

Drug	Dose	Sample Size (n_{ij})	i (drug index)	j (dose index)
	1	100	1	1
A	2	100	1	2
	3	50	1	3
	1	100	2	1
B	2	100	2	2
	3	50	2	3

$$\mu_A = \frac{100(\mu+\alpha_1+\beta_1+\gamma_{11})+100(\mu+\alpha_1+\beta_2+\gamma_{12})+50(\mu+\alpha_1)}{250}$$

$$= \frac{250\mu+250\alpha_1+100\gamma_{11}+100\gamma_{12}+100\beta_1+100\beta_2}{250}$$

$$= \mu+\alpha_1+\frac{2}{5}(\beta_1+\beta_2+\gamma_{11}+\gamma_{12})$$

$$\mu_B = \frac{100(\mu+\beta_1)+100(\mu+\beta_2)+50(\mu)}{250}$$

$$= \frac{250\mu+100(\beta_1+\beta_2)}{250}$$

$$= \mu+\frac{2}{5}(\beta_1+\beta_2)$$

H_0: $\mu_A = \mu_B \Rightarrow \mu+\alpha_1+\frac{2}{5}(\beta_1+\beta_2+\gamma_{11}+\gamma_{12}) = \mu+\frac{2}{5}(\beta_1+\beta_2)$

$$\Rightarrow \alpha_1+\frac{2}{5}(\gamma_{11}+\gamma_{12}) = 0$$

$$C = \begin{bmatrix} 0 & 1 & 0 & 0 & \frac{2}{5} & \frac{2}{5} \end{bmatrix}$$

df 1, 494

5. (15pts) Still using the data of Problem 4 (with balanced design of 100 samples in each cell). Now we introduce another interval variable "age" and the interaction between drug and dose, fit a model using the following SAS code

```
proc glm;
class drug;
model LDL= age dose drug drug*dose/ solution;
run;
```

and obtained the following output.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	86791.9439	21697.9860	60.26	<.0001
Error	595	214247.6617	360.0801		
Corrected Total	599	301039.6056			

R-Square	Coeff Var	Root MSE	LDL Mean
0.288307	15.19667	18.97578	124.8680

Source	DF	Type I SS	Mean Square	F Value	Pr > F
age	1	21309.98280	21309.98280	59.18	<.0001
dose	1	6664.73548	6664.73548	18.51	<.0001
drug	1	58218.74750	58218.74750	161.68	<.0001
dose*drug	1	598.47814	598.47814	1.66	0.1978

Source	DF	Type III SS	Mean Square	F Value	Pr > F
age	1	19205.20980	19205.20980	53.34	<.0001
dose	1	6683.65025	6683.65025	18.56	<.0001
drug	1	4691.53105	4691.53105	13.03	0.0003
dose*drug	1	598.47814	598.47814	1.66	0.1978

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	104.9204188	B	3.94321568	26.61	<.0001
age	0.4855570		0.06648601	7.30	<.0001
dose	5.3125392	B	1.34185926	3.96	<.0001
drug 0	-14.8100813	B	4.10298291	-3.61	0.0003
drug 1	0.0000000	B	.	.	.
dose*drug 0	-2.4479126	B	1.89876546	-1.29	0.1978
dose*drug 1	0.0000000	B	.	.	.

plug in the values of β_i's?

(a) (3pts) Write down the fitted model based on the above output. Is dose treated as categorical or interval variable?

$LDL = \beta_0 + \beta_1 \times age + \beta_2 \times dose + \beta_3 \times drug + \beta_4 \times drug \times dose + \varepsilon$

Dose was treated as continuous here since only drug was used in the class statement.

(b) (2pts) Why is the regression coefficient estimate for "drug 1" is 0 without estimate for standard error? Note the numerical value of drug is 0 for drug A and 1 for drug B.

Because drug 1 (drug B) was used as the reference group and embedded in the intercept.

(c) (3pts) Briefly explain what is the difference between Type I SS and Type III SS. Why the Type I SS of age is larger than the Type III SS of age, but the Type I SS of dose*drug is the same as the Type III SS of dose*drug?

Type I SS are from added-in-order tests, and they are mutually exclusive and together exhaustive pieces of the model SS. The sizes of Type I SS for a covariate depends on the order the covariate is added to the model, except when all predictors are uncorrelated. Type III SS are from added-last tests, and they are SS for each variable if it was entered last in the model. The size of Type III SS tells how much variance being explained by this variable after accounting for all other variables. Here Age was added first in the model, so its Type I SS is much larger than its Type III SS.

The variable added last into the model in added-in-order test is equivalent to the added-last test of this variable since SSs from these two tests are SS explained by this variable beyond the other variables. This is the reason why for dose*drug, the Type I SS is the same as the Type III SS.

(d) (4pts) Write the contrast matrix to estimate the average LDL level when drug A is used for an individual of age 40. Similarly, Write the contrast matrix to estimate the average LDL level when drug B is used for an individual of age 40.

$LDL = \beta_0 + \beta_1 age + \beta_2 dose + \beta_3 drug + \beta_4 drug\text{-}dose + \varepsilon$

drug A for individual 40:

$$[1 \quad 40 \quad \overline{dose} \quad 1 \quad \overline{dose}]$$

drug A: drug 0

drug B: drug 1

drug B for individual 40: because drug B was the reference.

$$[1 \quad 40 \quad \overline{dose} \quad 0 \quad 0]$$

where $\overline{dose}$ = grand mean of the dose variable

(e) (3pts) Write the contrast matrix to test the hypothesis that the average LDL level for the individuals of age 40 taking drug A is different from the average LDL level for the individuals of age 40 taking drug B. Write the formula to calculate the test-statistic and what is the degree of freedom of this test?

$H_0: \mu_1 = \mu_2$

$\theta = \mu_1 - \mu_2 = 0$

$\theta = \mu_1 - \mu_2 = \beta_0 + \beta_1(40) + \beta_2(\overline{dose}) + \beta_3(1) + \beta_4(\overline{dose}) - [\beta_0 + \beta_1(40) + \beta_2(\overline{dose}) + \beta_3(0) + \beta_4(0)]$

$= \beta_3 + \beta_4(\overline{dose}) = [0 \ 0 \ 0 \ 1 \ \overline{dose}] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = C\beta$

$C = [0 \ 0 \ 0 \ 1 \ \overline{dose}]$

$\hat{\theta} - \theta = C\hat{\beta} - 0 = C\hat{\beta}$

$Var(\hat{\theta}) = M\sigma^2$

$Var(\hat{\theta}) = C\,Var(\hat{\beta})\,C'$

$M = \frac{C\,Var(\hat{\beta})\,C'}{\sigma^2}$

$$F_{obs} = \frac{(\hat{\theta} - \theta)' M^{-1} (\hat{\theta} - \theta)/1}{MSE} = \frac{(C\hat{\beta})^2 / [Var(\hat{\theta})/\sigma^2]}{MSE}$$

$$= \frac{(C\hat{\beta})^2 / [C\,Var(\hat{\beta})\,C'/\sigma^2]}{360.080} = \frac{(C\hat{\beta})^2}{C\,Var(\hat{\beta})\,C'}$$

df: 1, 595