# Swarthmore Honors Exam 2014: Statistics

John W. Emerson, Yale University

**NAME:** ________________________________________

## Instructions:

This is a closed-book three-hour exam having 7 questions. You may not refer to notes or textbooks. You may use a calculator that does not do algebra or calculus. Normal, $t$, and $\chi^2$ tables should be supplied with this exam. Please note:

- A few of the questions have a considerable amount of description and background which is intended to help you. Read this material carefully and completely before working on the problem.
- Most questions have multiple parts. Number the questions and parts clearly in your work, and start each of the questions on a new page.
- Questions that explicitly ask for (or imply the need for) discussion are a chance to demonstrate your understanding of the material. As a guideline, a short paragraph is likely appropriate and preferable to either a single sentence or a longer essay.

**1.** A market research company employs a large number of typists to enter data into a computer. The time taken for new typists to learn the computer system is well-approximated by a normal distribution with a mean of 90 minutes and a standard deviation of 20 minutes.

a. Calculate the proportion of new typists that take more than two hours (120 minutes) to learn the computer system.

b. Calculate the time below which which 25% of new typists take to learn the computer system.

c. Two typists start learning the computer system at the same time. What is the probability they both learn the system before 70 minutes have passed? Explain any assumptions required for your calculation.

**2.** Nobel Laureate Linus Pauling (1901-1994) conducted a randomized experiment to study whether taking vitamin C supplements helps prevent the common cold. The results were reported in the *Proceedings of the National Academy of Sciences.* He randomly assigned 279 French skiers to two groups, group C (that took vitamin C supplements) and group S (that took a sugar pill placebo). Here are the results:

```
> pauling

        Caught a cold Did not catch a cold
Group C            17                  122
Group S            31                  109
```

The ultimate question: is there evidence that Vitamin C helps reduce the incidence rate of colds in this population? Let $p_c$ denote the (unknown) population incidence rate of colds for people taking vitamin C supplements, and let $p_s$ denote the (unknown) population incidence rate of colds for people taking the sugar placebo.

a. State the null ($H_0$) and alternative ($H_A$) hypotheses for an appropriate test.

b. What test statistic will you use and what is its (approximate) sampling distribution, assuming your null hypothesis $H_0$ is true? Explain your assumptions, and draw a rough picture of the sampling distribution (but clearly label the picture). Please carefully define any notation you introduce.

c. Find or approximate the p-value of the test and state and justify your conclusions.

**3.** The total lifetime in days of a certain very delicate mechanical component of a machine is known to be approximately $N(\mu = 100, \sigma^2 = 100)$. After 95 days, your component is still working. How much longer do you expect the component to work?

**4.** Suppose that $X_1$ and $X_2$ are independent from the $N(0, \theta)$ distribution (here $\theta$ is the variance), where $0 < \theta < \infty$ is unknown.

a. Find the maximum likelihood estimator of the variance $\theta$.

b. You are told to conduct a hypothesis test of $H_0 : \theta = 1$ versus the alternative $H_A : \theta > 1$. A sample of size $n = 2$ yields the MLE $\widehat{\theta} = 2.12$. What do you conclude, and why?

c. A student recommends using "an unbiased estimator of $\theta$, the sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

And when $H_0 : \theta = 1$ with a sample of size $n = 2$ the sampling distribution of $s^2$ is known to be $\chi^2_{1\ d.f.}$ and we can use this for conducting our hypothesis test." What do you think of this proposal compared to your solution in part (b), and why?

**5.** Suppose $X_1$, $X_1$, ..., $X_{n-1}$ are independent, identically distributed $N(\mu, \sigma^2)$ for fixed, unknown parameters $\mu$ and $\sigma^2$. Assume $X_n$ is independent $N(\tau, \sigma^2)$ where $\tau$ is very very large (with respect to $\mu$) so that $X_n$ acts as an outlier. It's so large, in fact, that you may treat $X_n$ as a constant taking the value $\tau$. Explore and describe the statistical properties of the standard t-test statistic and/or 95% confidence interval based on the full sample of size $n$ for conducting inference on $\mu$ in the presence of such an outlier.

**6.** Consider an independent, identically distributed set of random variables $X_1$, ..., $X_n$ that are known to be uniform on the interval $[0, 1]$. Let $\overline{X}$ denote the sample mean, and $\widetilde{X}$ denote the sample median. You are familiar with the Central Limit Theorem for the sample mean $\overline{X}$. However, there is also a Central Limit Theorem for the sample median $\widetilde{X}$.

Under certain conditions (satisfied here for the median of the $X_i$), stated casually:

$$\widetilde{X} \sim N\left(0.5, \sigma^2_{\widetilde{X}} = \frac{1}{4n}\right),$$

approximately, for large enough $n$. Or in general, if $f(m) > 0$, $F(m) = 1/2$, and $F$ is differentiable at $m$ then

$$\sqrt{n}(\widehat{m} - m) \to_d N\left(0, \frac{1}{[2f(m)]^2}\right)$$

where $m$ is the population median, $\widehat{m}$ is the sample median from a sample of size $n$, and $f$ and $F$ are the population density and cumulative distribution functions, respectively. The conditions essentially ensure the uniqueness of the median.

Now suppose that $Y_1$, ..., $Y_n$ are independent from the uniform distribution on the set $[-2, -1] \cup [1, 2]$. In this case the conditions above do not apply to the median of the $Y_i$. You should attempt parts (a) through (e) on this exam. You may choose to answer parts (f) through (j); if you choose not to do so, please come to the oral exam prepared to discuss them.

a. Show that the variance of $X_1$ is $1/12$.

b. What is $\mathbb{E}(X_1^2)$?

c. What is $\mathbb{E}\left((1 + X_1)^2\right)$?

d. Find the variance of $Y_1$.

e. Suppose $n = 100$. What is $P(\overline{X} < 0.45)$, approximately?

f. Suppose $n = 100$. What is $P(\widetilde{X} < 0.45)$, approximately?

g. Suppose $n = 100$. Can you find $P(\overline{Y} < -0.05)$, approximately? If so, do it.

h. Suppose $n = 100$. Can you find $P(\widetilde{Y} < -0.05)$, approximately? If so, do it.

i. Suppose $n = 100$. Can you find $P(\overline{Y} < -1.05)$, approximately? If so, do it.

j. Suppose $n = 100$. Can you find $P(\widetilde{Y} < -1.05)$, approximately? If so, do it.

**NOTE:** Question 7 begins on the following page and is the last question on the exam. So all material from here to the end pertains to question 7, with the actual questions posed appearing on page 14.

**7.** This problem examines the results of a study of force required to move fingers on a cadaver's hand that has its tendons attached to a special testing device. The medical researchers were concerned that the use of special wraps (commonly used as part of physical therapy) and the presence of swelling (common after injuries) could result in increased force required to move fingers (i.e. during physical therapy exercises) and could thus increase the risk of re-injury. The study involved repeatedly measuring the force required to move each of the digits under the various experimental conditions. It was conducted over several weeks, repeatedly freezing and thawing the hand in an attempt to provide an even-handed (sorry about that) experimental design across sequences of measurements.

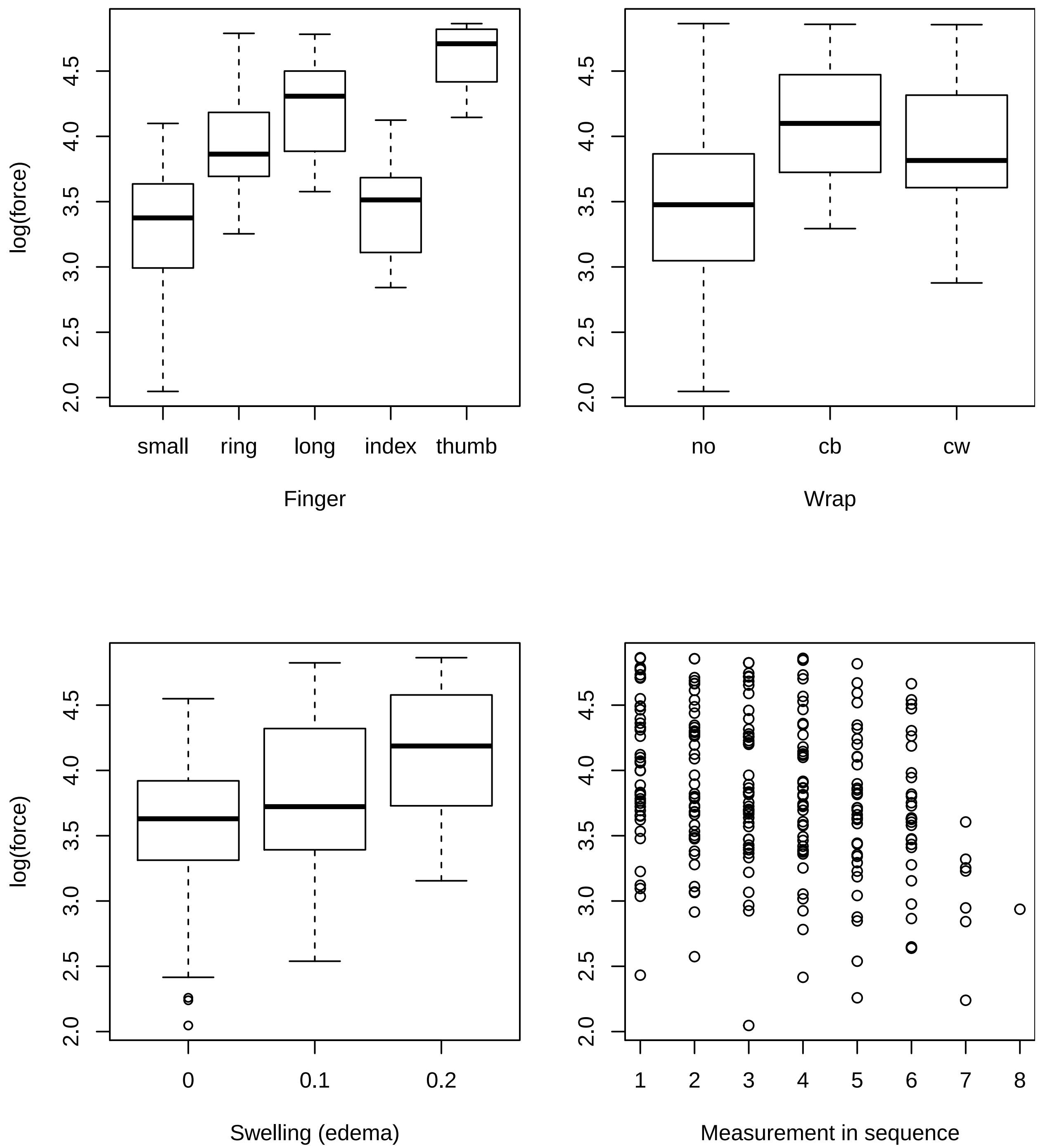
You will be presented with some basic exploratory data analysis, two linear models, their regression and analysis of variance summaries, an array of supporting plots, and so on. There are a few focused questions and a few open-response questions posed on page 14. Some notes:

- The raw response, `force`, is log-transformed and called `logforce`.
- The categorical variable `finger.f` has obvious labels.
- The categorical variable `wrap.f` has three levels, one for "no wrap used" and two for specific brands or styles of wrap denoted "cb" and "cw".
- The categorical variable `edema.f` refers to the degree of swelling induced (via the injection of saline). Explorations involving this as a quantitative variable were not satisfactory.
- The variable `measure` records the number in sequence of the particular measurement. The researchers were aware that the first movement of a digit might require higher force (and thus created a categorical variable `firstmeasure`), and that successive movements might require successively lower force as the digit/tendon warms up or becomes better lubricated. It is not known exactly why sequences are of different length, but the induced swelling is very short-lived (saline leaks out) and proper attachment of the tendons to the measuring device can be difficult. For the purpose of this problem you may disregard this particular quirk of the study.

The first four and last three sets of measurements are provided to help you better understand the nature of the study design:

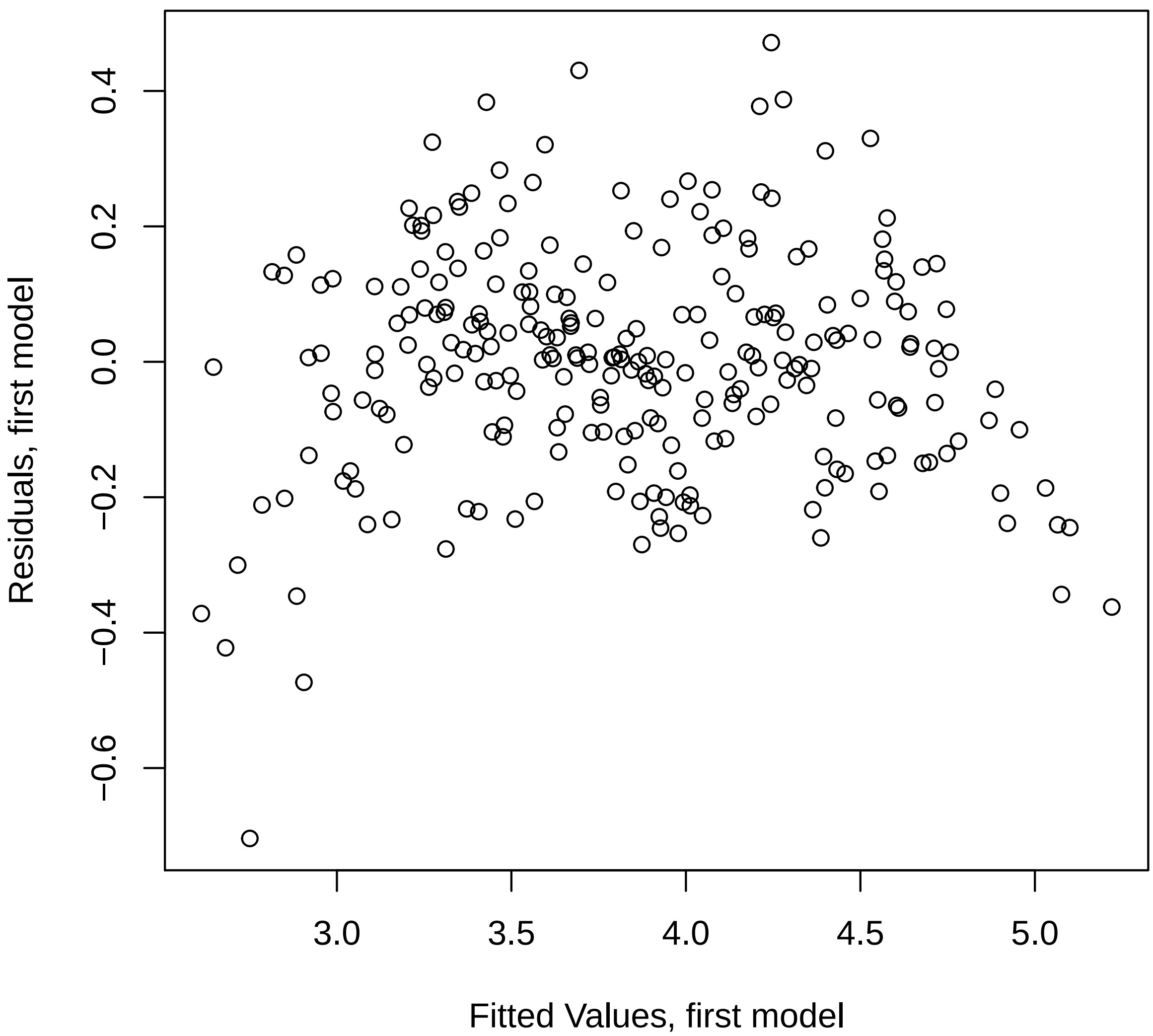
```
> head(y,23)

   measure     force logforce edema.f finger.f wrap.f firstmeasure
1        1 11.384098 2.432218       0    small     no         TRUE
2        2 13.117214 2.573925       0    small     no        FALSE
3        3  7.741917 2.046649       0    small     no        FALSE
4        4 11.196633 2.415613       0    small     no        FALSE
5        5  9.570255 2.258660       0    small     no        FALSE
6        6 13.992317 2.638508       0    small     no        FALSE
7        7  9.390852 2.239736       0    small     no        FALSE
8        1 32.400901 3.478186       0    small     cb         TRUE
9        2 28.709715 3.357236       0    small     cb        FALSE
10       3 27.989780 3.331839       0    small     cb        FALSE
11       4 30.547967 3.419298       0    small     cb        FALSE
12       5 26.939468 3.293592       0    small     cb        FALSE
13       1 25.175020 3.225852       0    small     cw         TRUE
14       2 21.434793 3.065015       0    small     cw        FALSE
15       3 25.013708 3.219424       0    small     cw        FALSE
16       4 20.428243 3.016918       0    small     cw        FALSE
17       5 17.768559 2.877431       0    small     cw        FALSE
18       1 22.665163 3.120829     0.1    small     no         TRUE
19       2 18.458636 2.915532     0.1    small     no        FALSE
20       3 19.434858 2.967068     0.1    small     no        FALSE
21       4 16.143180 2.781498     0.1    small     no        FALSE
22       5 12.665357 2.538871     0.1    small     no        FALSE
23       6 14.132780 2.648497     0.1    small     no        FALSE
```

```
> tail(y, 12)

    measure    force logforce edema.f finger.f wrap.f firstmeasure
241       1 129.4734 4.863476     0.2    thumb     no         TRUE
242       2 108.5755 4.687446     0.2    thumb     no        FALSE
243       3 114.9582 4.744569     0.2    thumb     no        FALSE
244       4 128.8510 4.858657     0.2    thumb     no        FALSE
245       1 128.7879 4.858166     0.2    thumb     cb         TRUE
246       2 128.4033 4.855176     0.2    thumb     cb        FALSE
247       3 124.5335 4.824575     0.2    thumb     cb        FALSE
248       4 126.9949 4.844147     0.2    thumb     cb        FALSE
249       1 113.5756 4.732469     0.2    thumb     cw         TRUE
250       2 128.4493 4.855534     0.2    thumb     cw        FALSE
251       3 108.0423 4.682523     0.2    thumb     cw        FALSE
252       4 127.2032 4.845785     0.2    thumb     cw        FALSE
```


log(force)
2.0
2.5
3.0
3.5
4.0
4.5
small
ring
long
index
thumb
Finger
no
cb
cw
Wrap
0
0.1
0.2
Swelling (edema)
1
2
3
4
5
6
7
8
Measurement in sequence

The first linear model for your consideration:

```
> lm.finger <- lm(logforce ~ firstmeasure + measure +
+                    finger.f + edema.f + wrap.f, data=y)
> summary(lm.finger)

Residuals:
     Min       1Q   Median       3Q      Max
-0.70394 -0.10217  0.00614  0.10117  0.47142

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.854830   0.045678  62.499  < 2e-16
firstmeasureTRUE  0.085550   0.036365   2.353   0.0194
measure          -0.034747   0.008126  -4.276 2.75e-05
finger.fring      0.646304   0.032917  19.634  < 2e-16
finger.flong      0.938229   0.033362  28.123  < 2e-16
finger.findex     0.202804   0.032582   6.224 2.13e-09
finger.fthumb     1.290097   0.037217  34.664  < 2e-16
edema.f0.1        0.203872   0.026079   7.817 1.67e-13
edema.f0.2        0.522792   0.026217  19.941  < 2e-16
wrap.fcb          0.501842   0.026093  19.233  < 2e-16
wrap.fcw          0.357629   0.026065  13.721  < 2e-16
---
Residual standard error: 0.1692 on 241 degrees of freedom
Multiple R-squared:  0.9221,  Adjusted R-squared:  0.9189
F-statistic: 285.4 on 10 and 241 DF,  p-value: < 2.2e-16

> anova(lm.finger)

Analysis of Variance Table

Response: logforce
              Df Sum Sq Mean Sq F value    Pr(>F)
firstmeasure   1  2.457  2.4568  85.828 < 2.2e-16
measure        1  4.419  4.4195 154.396 < 2.2e-16
finger.f       4 51.125 12.7813 446.521 < 2.2e-16
edema.f        2 12.387  6.1937 216.381 < 2.2e-16
wrap.f         2 11.303  5.6515 197.438 < 2.2e-16
Residuals    241  6.898  0.0286
---
```


0.4
0.2
0.0
−0.2
−0.4
−0.6
Residuals, first model
3.0
3.5
4.0
4.5
5.0
Fitted Values, first model

Here are the cases in the data corresponding to the extreme fitted values, to aid your exploration:

```
> y[lm.finger$fitted < 2.8,]

  measure     force logforce edema.f finger.f wrap.f firstmeasure
2       2 13.117214 2.573925       0    small     no        FALSE
3       3  7.741917 2.046649       0    small     no        FALSE
4       4 11.196633 2.415613       0    small     no        FALSE
5       5  9.570255 2.258660       0    small     no        FALSE
6       6 13.992317 2.638508       0    small     no        FALSE
7       7  9.390852 2.239736       0    small     no        FALSE
     residuals fittedvalues
2 -0.211410820     2.785336
3 -0.703939860     2.750589
4 -0.300228996     2.715842
5 -0.422435148     2.681095
6 -0.007839597     2.646348
7 -0.371864886     2.611601

> y[lm.finger$fitted > 4.9,]

    measure    force logforce edema.f finger.f wrap.f firstmeasure
231       1 110.7688 4.707445     0.1    thumb     cb         TRUE
245       1 128.7879 4.858166     0.2    thumb     cb         TRUE
246       2 128.4033 4.855176     0.2    thumb     cb        FALSE
247       3 124.5335 4.824575     0.2    thumb     cb        FALSE
248       4 126.9949 4.844147     0.2    thumb     cb        FALSE
249       1 113.5756 4.732469     0.2    thumb     cw         TRUE
250       2 128.4493 4.855534     0.2    thumb     cw        FALSE
251       3 108.0423 4.682523     0.2    thumb     cw        FALSE
     residuals fittedvalues
231 -0.1939992     4.901444
245 -0.3621982     5.220365
246 -0.2448908     5.100067
247 -0.2407452     5.065320
248 -0.1864258     5.030573
249 -0.3436829     5.076152
250 -0.1003204     4.955854
251 -0.2385846     4.921107
```

Here I present the second linear model for your consideration, with some interactions between some of the categorical variables:

```
> lm.fingeri <- lm(logforce ~ firstmeasure + measure +
+                    finger.f * edema.f * wrap.f, data=y)
> summary(lm.fingeri)

Residuals:
     Min       1Q   Median       3Q      Max
-0.33894 -0.05801 -0.00294  0.04392  0.33576

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                        2.468440   0.043186  57.158  < 2e-16
firstmeasureTRUE                   0.099468   0.021378   4.653 5.87e-06
measure                           -0.027616   0.004856  -5.687 4.42e-08
finger.fring                       1.047859   0.052956  19.787  < 2e-16
finger.flong                       1.296901   0.052956  24.490  < 2e-16
finger.findex                      0.651224   0.052956  12.298  < 2e-16
finger.fthumb                      1.840611   0.062392  29.501  < 2e-16
edema.f0.1                         0.440353   0.055159   7.983 1.00e-13
edema.f0.2                         1.073544   0.058160  18.458  < 2e-16
wrap.fcb                           0.970544   0.058160  16.687  < 2e-16
wrap.fcw                           0.675442   0.058160  11.613  < 2e-16
finger.fring:edema.f0.1           -0.326166   0.077949  -4.184 4.24e-05
finger.flong:edema.f0.1           -0.345258   0.080046  -4.313 2.50e-05
finger.findex:edema.f0.1          -0.501283   0.075405  -6.648 2.64e-10
finger.fthumb:edema.f0.1          -0.029443   0.093622  -0.314 0.753466
finger.fring:edema.f0.2           -0.496167   0.080046  -6.199 3.08e-09
finger.flong:edema.f0.2           -0.362341   0.080046  -4.527 1.02e-05
finger.findex:edema.f0.2          -0.768072   0.078657  -9.765  < 2e-16
finger.fthumb:edema.f0.2          -0.549885   0.091050  -6.039 7.17e-09
finger.fring:wrap.fcb             -0.533695   0.080046  -6.667 2.37e-10
finger.flong:wrap.fcb             -0.467347   0.082039  -5.697 4.20e-08
finger.findex:wrap.fcb            -0.364301   0.080046  -4.551 9.13e-06
finger.fthumb:wrap.fcb            -0.796232   0.095387  -8.347 1.03e-14
finger.fring:wrap.fcw             -0.371329   0.080046  -4.639 6.24e-06
finger.flong:wrap.fcw             -0.423363   0.080046  -5.289 3.14e-07
finger.findex:wrap.fcw            -0.186050   0.080046  -2.324 0.021089
finger.fthumb:wrap.fcw            -0.691932   0.091050  -7.599 1.04e-12
edema.f0.1:wrap.fcb               -0.332950   0.083478  -3.988 9.25e-05
edema.f0.2:wrap.fcb               -0.566555   0.083682  -6.770 1.33e-10
edema.f0.1:wrap.fcw               -0.098964   0.080268  -1.233 0.219019
edema.f0.2:wrap.fcw               -0.373396   0.085491  -4.368 1.99e-05
finger.fring:edema.f0.1:wrap.fcb   0.484272   0.115212   4.203 3.93e-05
```

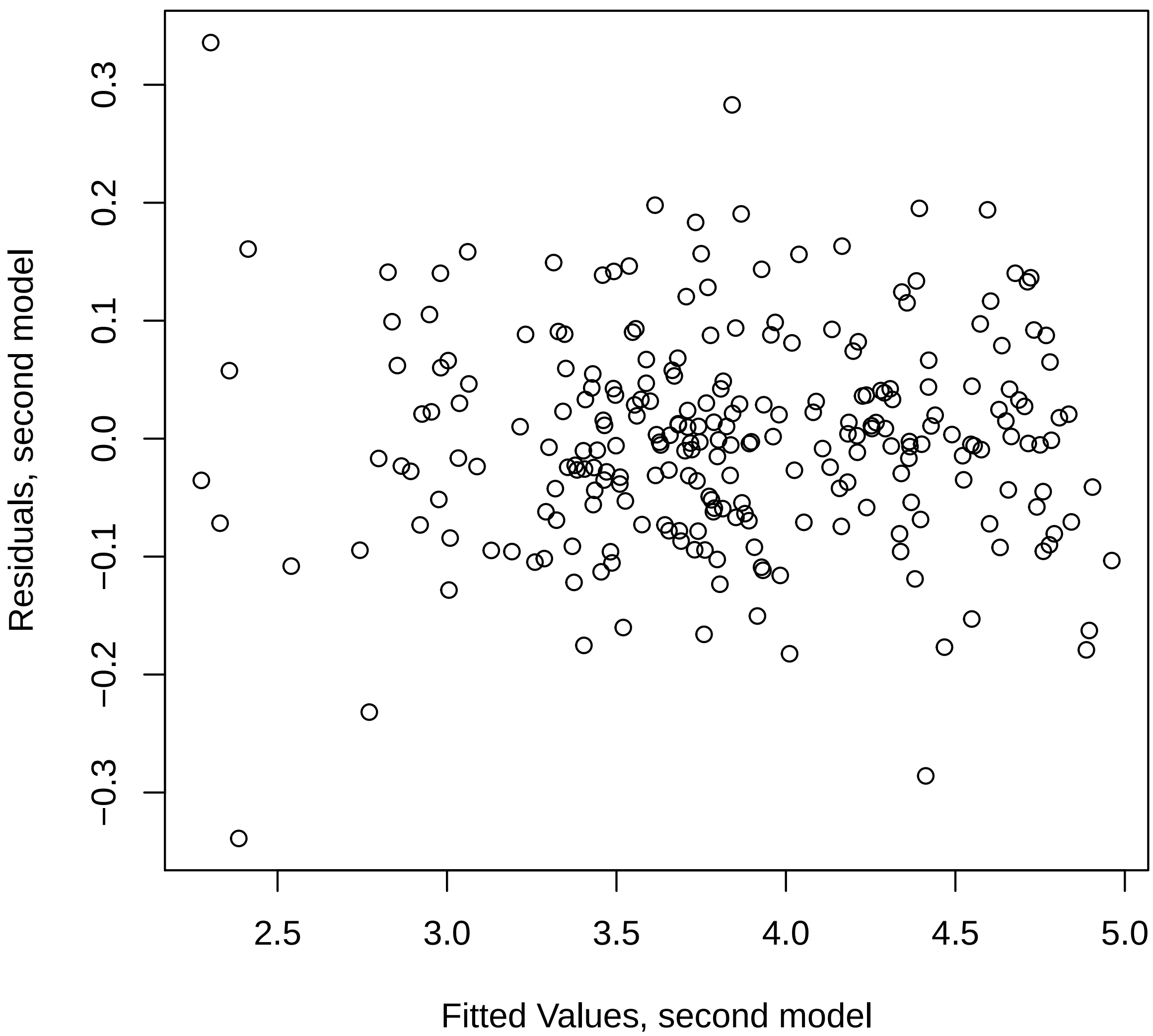
```
finger.flong:edema.f0.1:wrap.fcb   0.387371   0.118088   3.280 0.001218
finger.findex:edema.f0.1:wrap.fcb  0.564464   0.113506   4.973 1.39e-06
finger.fthumb:edema.f0.1:wrap.fcb  0.253348   0.132771   1.908 0.057769
finger.fring:edema.f0.2:wrap.fcb   0.559110   0.115284   4.850 2.44e-06
finger.flong:edema.f0.2:wrap.fcb   0.298238   0.116641   2.557 0.011284
finger.findex:edema.f0.2:wrap.fcb  0.391592   0.114362   3.424 0.000745
finger.fthumb:edema.f0.2:wrap.fcb  0.449223   0.132739   3.384 0.000855
finger.fring:edema.f0.1:wrap.fcw   0.072125   0.112829   0.639 0.523378
finger.flong:edema.f0.1:wrap.fcw   0.438972   0.115692   3.794 0.000195
finger.findex:edema.f0.1:wrap.fcw  0.143650   0.111249   1.291 0.198072
finger.fthumb:edema.f0.1:wrap.fcw  0.165801   0.130614   1.269 0.205739
finger.fring:edema.f0.2:wrap.fcw   0.324337   0.116641   2.781 0.005931
finger.flong:edema.f0.2:wrap.fcw   0.331797   0.116641   2.845 0.004898
finger.findex:edema.f0.2:wrap.fcw  0.254920   0.115692   2.203 0.028678
finger.fthumb:edema.f0.2:wrap.fcw  0.380427   0.130858   2.907 0.004048
---
Residual standard error: 0.09907 on 205 degrees of freedom
Multiple R-squared:  0.9773,  Adjusted R-squared:  0.9722
F-statistic: 191.8 on 46 and 205 DF,  p-value: < 2.2e-16
```

*> anova(lm.fingeri)*

```
Analysis of Variance Table

Response: logforce
                         Df Sum Sq Mean Sq   F value    Pr(>F)
firstmeasure              1  2.457  2.4568  250.3041 < 2.2e-16
measure                   1  4.419  4.4195  450.2735 < 2.2e-16
finger.f                  4 51.125 12.7813 1302.2113 < 2.2e-16
edema.f                   2 12.387  6.1937  631.0417 < 2.2e-16
wrap.f                    2 11.303  5.6515  575.7977 < 2.2e-16
finger.f:edema.f          8  1.905  0.2382   24.2667 < 2.2e-16
finger.f:wrap.f           8  1.758  0.2197   22.3854 < 2.2e-16
edema.f:wrap.f            4  0.555  0.1386   14.1260 3.371e-10
finger.f:edema.f:wrap.f  16  0.669  0.0418    4.2574 3.895e-07
Residuals               205  2.012  0.0098
---
```

*> BIC(lm.fingeri, lm.finger)*

```
           df       BIC
lm.fingeri 48 -236.6663
lm.finger  12 -125.2317
```


Residuals, second model
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
2.5
3.0
3.5
4.0
4.5
5.0
Fitted Values, second model

**Questions posed for this problem 7:**

a. **Using the first model,** interpret the coefficient listed in the regression output as `firstmeasureTRUE` on page 8.

b. **Using the first model,** interpret the coefficient listed in the regression output as `measure` on page 8.

c. **Using the first model presented on page 8,** what is the predicted `logforce` for a hypothetical future $4^{th}$ measurement of the thumb that has no swelling (edema) but is wrapped in Cobain (cb) wrap?

d. **Using the first model,** examine the regression output, the residual plot provided, and the investigation of selected points presented on pages 8-10. Comment on anything that worries you here, and explain why.

e. Now consider both models. Which do you prefer, and why? On this exam, please be brief. Come to the oral exam prepared to discuss this problem further, and I'll have the data set available for "live" exploration.