March 3, 2008

Some Solutions to Homework 3

1. (a) $\log(ab) + \log(c+d) = \log(a) + \log(b) + \log(c+d)$
   (b) $\frac{1}{e^x + e^y}$
   (c) $\frac{1}{\sqrt{w}}$
   (d) $(x-y)^3 + 3x^2y - 3xy^2$

2. $(b_1 + b_5 x_2)(g - f) + b_2(g^2 - f^2) + b_3(g^3 - f^3)$

3. (a) Two-sample (unpaired) $t$-test; Wilcoxon-Mann-Whitney two-sample rank-sum test
   (b) Analysis of variance; Kruskal-Wallis nonparametric ANOVA
   (c) Pearson linear correlation test; linear regression; Spearman nonparametric correlation test
   (d) $\chi^2$ test for $2 \times 2$ table; Fisher's "exact" test

4. Prediction plays an important though frequently overlooked role. In testing treatments on a sample of patients we are much more interested in using the resulting data to predict the likely benefit of a treatment on **future** patients.

```
x ← 1:5
y ← c(98, 198, 315, 380, 530)

Lxx ← sum((x - mean(x))^2)
Lxy ← sum((x - mean(x))*(y - mean(y)))
b ← Lxy / Lxx
a ← mean(y) - b * mean(x)
c(a=a, b=b)

a     b
 -9.6 104.6
```

5. 
```
yhat ← a + b * x
resid ← y - yhat
rbind(x, y, yhat, resid)

      [,1]   [,2]  [,3]   [,4]  [,5]
    x    1    2.0   3.0    4.0   5.0
    y   98  198.0 315.0  380.0 530.0
 yhat   95  199.6 304.2  408.8 513.4
resid    3   -1.6  10.8  -28.8  16.6
```

6. 
```
sse ← function(x, y, a, b) {
  yhat ← a + b * x
  sum((y - yhat)^2)
}

> sse(x, y, a, b)         # Could also do:
[1] 1233.2                # for(ac in c(-10,10))
> sse(x, y, a-10, b*.9)   #   for(bf in c(.9,1/.9))
[1] 10888.84              #     print(sse(x,y,a+ac,b*bf))
> sse(x, y, a+10, b*.9)
[1] 4612.838
> sse(x, y, a-10, b/.9)
[1] 5675.716
> sse(x, y, a+10, b/.9)
[1] 12649.05
```

   SSE is lowest when evaluated at the least squares estimates, as it should be. These estimates were derived from formulas that minimize SSE.

Some Solutions to the Midterm Exam

1. (a) 0.45
   (b) 0.05
   (c) 0.40
   (d) Bimodal and skewed (to the right or postively)

2. (a) $H_0 : \mu_1 = \mu_2$; $H_1 : \mu_1 \neq \mu_2$
   (b) $s2_{\text{pooled}} = 26.14; T = 0.64; p\text{-value} = 0.53$
   (c) Fail to reject $H_0$. Something like "we do not have enough evidence to show that the treatment and placebo groups have different tumor volume". It would be wrong to indicate that tumor volume in treatment and placebo are "the same" or "equal" (abscence of evidence is not evidence of abscence)
   (d) $[-24.1, 38.1]$
   (e) Normality and equal variance in the two groups. To verify, several answers from talking with the investigators before conducting the experiment to using simple desciptive plots (dot plots, box plots), or something similar would be great. Based on the given $s_1$ and $s_2$, the equal variances assumption may not hold.

3. (a) Narrower; If $\alpha$ increases, the $t_{\text{crit}}$ (also called $t'$ in the book) decreases. In other words, a 90% CI ($\alpha = 0.10$) is narrower than a 95% CI ($\alpha = 0.05$)
   (b) Narrower; Smaller standard error of the mean
   (c) Wider; larger standard error of the mean
   (d) Narrower; Can use the Normal distribution rather than the t distribution; $Z_{\text{crit}}$ is smaller than $t_{\text{crit}}$, especially for $n = 10$
   (e) 3.5 Unchanged; Just shifts the confidence interval to the right, so the width is unchanged

4. 
```
x ← c(36, 20, 3, 3, 2, 30, 0, 10, 22)/10
y ← c(1700,3078,1820,2706,2086,2299,676,2088,2013)
```

   1-6 Use the `Rcmdr` menus or the R code below

```
Lxx ← sum((x - mean(x))^2)
Lxy ← sum((x - mean(x)) * (y - mean(y)))
b ← Lxy/Lxx
a ← mean(y) - b * mean(x)
SSE ← sum((y - a - b*x)^2)
n ← length(x)
s2y.x ← SSE/(n-2)
SST ← sum((y - mean(y))^2)
R2 ← (SST - SSE) / SST

sy.x ← sqrt(s2y.x)
se.b ← sy.x / sqrt(Lxx)
tstat ← b / se.b
pval ← 2*(1-pt(abs(tstat),n-2))
data.frame(a, b, SSE, s2y.x, R2, sy.x, se.b, tstat, pval)
```

```
        a       b     SSE    s2y.x         R2     sy.x     se.b     tstat      pval
 1894.818 112.114 3435727 490818.2 0.04997965 700.5842 184.7485 0.6068469 0.5631094
```

   7 Use `b` $\pm 2\times$ `se.b` then use the correct $t$ critical value in place of the 2: `qt(.975, 902)` = 2.365

8
```
yhat ← a + b*3
se.Ey.x ← sy.x * sqrt(1/n + (3-mean(x))^2 / Lxx)
se.Ey.x

[1] 376.7138

yhat + c(-1, 1)*qt(.975, n-2)*se.Ey.x

[1] 1340.374 3121.947
```

9
```
se.yhat ← sy.x * sqrt(1 + 1/n + (3-mean(x))^2 / Lxx)
se.yhat

[1] 795.4442

yhat + c(-1, 1)*qt(.975, n-2)*se.yhat

[1]  350.2336 4112.0869
```

10 $s_{y.x}$

5. 1 Solutions below use the R `Design` package. Similar output is obtained with the `Rcmdr` menus and with the builtin R `lm` function.

```
f ← ols(maxfwt  ~ age + sex)
f

Linear Regression Model

ols(formula = maxfwt ~ age + sex)

  n Model L.R. d.f.     R2 Sigma
 99      54.58    2 0.4238 9.847

Residuals:
    Min     1Q Median    3Q  Max
 -34.39 -4.649  1.352 6.169 32.5

Coefficients:
             Value Std. Error t value   Pr(>|t|)
 Intercept  24.290     3.4682  7.0037 3.409e-010
       age   2.808     0.3362  8.3515 5.098e-013
sex=female  -1.523     2.0573 -0.7404 4.609e-001

Residual standard error: 9.847 on 96 degrees of freedom
Adjusted R-Squared: 0.4118

an ← anova(f)
print(an, 'names')

                Analysis of Variance          Response: maxfwt

 Factor     d.f. Partial SS         MS     F      P         Tested
 age         1   6762.44720 6762.44720 69.75 <.0001 age
 sex         1     53.14713   53.14713  0.55 0.4609 sex=female
 REGRESSION  2   6845.95126 3422.97563 35.30 <.0001 age,sex=female
 ERROR      96   9307.88713   96.95716

# Note: also get overall F from R-squared:
r2 ← 0.4238
(r2/2)/((1-r2)/96)
[1] 35.30441
```

The interpretation of coefficients was not requested, but here they are.

**Intercept** : estimate of mean `maxfwt` for a zero year-old male

**Coefficient of age** : estimate of the increase in mean `maxfwt` per one-year increase in age if adjusting for sex

**Coefficient of sex** : estimate of the difference in mean `maxfwt` between females and males (female mean - male mean) if age is held constant

2 The following is a detailed analysis when the interaction is not included (not requested but may help in understanding).

```
f ← ols(maxfwt ∼ age + sex + group)
f

Linear Regression Model

ols(formula = maxfwt ∼ age + sex + group)

  n Model L.R. d.f.     R2 Sigma
 99       63.9    4 0.4756 9.493

Residuals:
    Min     1Q Median    3Q   Max
 -34.11 -4.677  1.231 5.318 32.22

Coefficients:
                                                Value Std. Error t value
                                    Intercept  27.467      3.538  7.7632
                                          age   2.679      0.327  8.1922
                                    sex=female  -1.868      1.988 -0.9395
      group=blood lead >= 40mg/100ml in 1973  -7.648      2.510 -3.0472
group=blood lead >= 40 in 1972, < 40 in 1973  -1.691      2.658 -0.6361
                                                Pr(>|t|)
                                    Intercept 9.991e-012
                                          age 1.260e-012
                                    sex=female 3.499e-001
      group=blood lead >= 40mg/100ml in 1973 2.998e-003
group=blood lead >= 40 in 1972, < 40 in 1973 5.263e-001

Residual standard error: 9.493 on 94 degrees of freedom
Adjusted R-Squared: 0.4533

print(anova(f), 'dots')

                Analysis of Variance          Response: maxfwt

 Factor     d.f. Partial SS         MS     F      P Tested
 age         1   6047.92047 6047.92047 67.11 <.0001 .
 sex         1     79.54343   79.54343  0.88 0.3499  .
 group       2    836.85244  418.42622  4.64 0.0119   ..
 REGRESSION  4   7682.80370 1920.70093 21.31 <.0001 ....
 ERROR      94   8471.03468   90.11739

Subscripts correspond to:
[1] age
[2] sex=female
[3] group=blood lead >= 40mg/100ml in 1973
[4] group=blood lead >= 40 in 1972, < 40 in 1973
```

**Intercept** estimate of mean wrist-finger tapping score for a newborn male with blood lead $< 40$ mg in both years

**Age coefficient** : estimate of the increase in the mean `maxfwt` per year increase in age, holding sex and exposure constant.

**Sex coefficient** :estimate of difference in mean `maxfwt` between females and males (female - male), holding age and exposure constant.

**-7.6481** : estimates the mean `maxfwt` for children having blood lead level $\geq$ 40 mg in 1973 minus the mean `maxfwt` for children not exposed to $\geq$ 40 mg of lead, for children of the same age and sex.

**-1.6908** : estimates the mean `maxfwt` for children having blood lead level $\geq$ 40 mg in 1972 but not in 1973 minus the mean `maxfwt` for children not exposed to $\geq$ 40 mg of lead (in either year), for children of the same age and sex.

To estimate the mean difference in `maxfwt` for those exposed in 1972 but not in 1973 compared with those exposed in 1973 (and 1972), use -1.6908 - (-7.6481).

In the `anova` above, the first $F$ (67.11) tests $H_0$: there is no association between age and `maxfwt` after adjusting for sex and group.
Second $F$ (0.88) tests $H_0$: there is no association between sex and `maxfwt`, hold age and group constant.
Third $F$ (4.64) tests $H_0$: no association between degree of exposure and `maxfwt`, when comparing children of the same age and sex.

```
anova(ols(maxfwt ~ age))

Factor     d.f. Partial SS         MS     F      P
 age         1   6792.804   6792.80412 70.39 <.0001
 REGRESSION  1   6792.804   6792.80412 70.39 <.0001
 ERROR      97   9361.034     96.50551

anova(ols(maxfwt ~ sex))

 Factor     d.f.  Partial SS        MS   F      P
 sex         1      83.50406  83.50406 0.5 0.4794
 REGRESSION  1      83.50406  83.50406 0.5 0.4794
 ERROR      97   16070.33432 165.67355

anova(ols(maxfwt ~ group))

 Factor     d.f. Partial SS       MS    F      P
 group       2    1600.088  800.0442 5.28 0.0067
 REGRESSION  2    1600.088  800.0442 5.28 0.0067
 ERROR      96   14553.750  151.6016
```

These tests assess whether the individual variables are associated with `maxfwt`, without adjusting for any other variables. The partial $F$ statistics for age and group are smaller than unadjusted $F$ ratios, as is often the case, because some of the association between these variables and outcome is explained by the other two variables. In other words, the distribution of age and sex differs by group, and age and sex (especially age) are also related to `maxfwt`, detracting from the apparent unadjusted relationship between group and `maxfwt`.

Next the requested analysis:

```
f ← ols(maxfwt ~ age*sex + group)
f

Linear Regression Model

ols(formula = maxfwt ~ age * sex + group)

  n Model L.R. d.f.     R2 Sigma
 99      67.03    5 0.4919 9.395

Residuals:
    Min     1Q Median    3Q   Max
 -32.93 -5.369 0.8814 5.124 32.71

Coefficients:
                                                Value Std. Error t value
```

```
                                     Intercept  22.783     4.4294  5.1437
                                           age   3.156     0.4256  7.4158
                                    sex=female   9.980     7.1397  1.3979
          group=blood lead >= 40mg/100ml in 1973  -6.966     2.5151 -2.7697
     group=blood lead >= 40 in 1972, < 40 in 1973  -1.958     2.6350 -0.7429
                                 age * sex=female  -1.146     0.6639 -1.7263
                                                   Pr(>|t|)
                                     Intercept 1.490e-006
                                           age 5.523e-011
                                    sex=female 1.655e-001
          group=blood lead >= 40mg/100ml in 1973 6.774e-003
     group=blood lead >= 40 in 1972, < 40 in 1973 4.594e-001
                                 age * sex=female 8.762e-002

     Residual standard error: 9.395 on 93 degrees of freedom
     Adjusted R-Squared: 0.4646
```

The slope of age for females is 1.146 lower than the slope of age for males. The $t$ statistic of -1.7263 and $P = 0.088$ test $H_0$ : age effect is the same for both sexes. This can also be worded as the hypothesis that the effects of age and sex are additive.

3
```
attach(lead)
f ← ols(maxfwt ∼ ld72 + ld73)
f

Linear Regression Model

ols(formula = maxfwt ∼ ld72 + ld73)

  n Model L.R. d.f.     R2 Sigma
 99      10.55    2 0.1011  12.3

Residuals:
   Min     1Q Median    3Q   Max
 -42.2 -6.339   1.13 7.477 31.39

Coefficients:
             Value Std. Error  t value Pr(>|t|)
Intercept  64.28921      3.977  16.1665  0.00000
     ld72  -0.02476      0.101  -0.2451  0.80689
     ld73  -0.36298      0.170  -2.1358  0.03524

Residual standard error: 12.3 on 96 degrees of freedom
Adjusted R-Squared: 0.08233
```

**64.3** is the estimated mean maxfwt in the population when the lead levels in both years are zero.

**0.0248** is the decrease in estimated mean maxfwt per unit increase in the 1972 lead level, holding the 1973 lead level constant (if that's possible)

**0.363** is the decrease in estimated mean maxfwt per unit increase in the 1973 lead level, holding the 1972 lead level constant (if that's possible)

The best combination of lead levels for predicting maxfwt (either mean or raw values) is $0.0248 \times ld72 + 0.363 \times ld73$.

ld73 is significantly associated with maxfwt holding ld72 constant. The reverse is not true.

4
```
f ← ols(maxfwt ∼ ld72 + ld73 + age + sex)
```

The $R^2$ increased from 0.10 to 0.47, indicating that age and sex explain much more of the variation in maxfwt than did the two lead levels.

```
anova(f)
```

```
Analysis of Variance          Response: maxfwt

 Factor     d.f. Partial SS        MS     F      P
 ld72        1     10.7665    10.7665  0.12 0.7317
 ld73        1    287.2681   287.2681  3.15 0.0790
 age         1   5899.4523  5899.4523 64.78 <.0001
 sex         1     53.6431    53.6431  0.59 0.4447
 REGRESSION  4   7593.7308  1898.4327 20.85 <.0001
 ERROR      94   8560.1076    91.0650
```

```
anova(f, ld72, ld73)  # test joint contribution of ld72,ld73
                      # adjusted for age, sex
```

```
Factor     d.f. Partial SS       MS    F      P
 ld72        1     10.7665   10.7665 0.12 0.7317
 ld73        1    287.2681  287.2681 3.15 0.0790
 REGRESSION  2    747.7795  373.8898 4.11 0.0195
 ERROR      94   8560.1076   91.0650
```

The required $F$ is 4.11, P=.0195. This tests $H_0$ : at least one of the two lead levels is associated with maxfwt after controlling for age and sex. We have evidence that there is some association. You can also obtain this partial $F$-test by subtracting SSRs for the complete and reduced model, or by the "difference in $R^2$ test."

5
```
f ← ols(maxfwt ∼ age + sex + area + totyrs + ld72 + ld73)
f
```

```
  n Model L.R. d.f.     R2 Sigma
 99      67.17    7 0.4926  9.49

Residuals:
    Min     1Q Median    3Q   Max
 -31.56 -4.615  1.227 5.443 30.53

Coefficients:
                  Value Std. Error t value   Pr(>|t|)
    Intercept  29.35998    5.42158  5.4154 4.946e-007
          age   2.83831    0.40986  6.9251 6.014e-010
   sex=female  -1.27127    2.03164 -0.6257 5.331e-001
  area=1-2.5m   4.02071    2.22118  1.8102 7.357e-002
area=2.5-4.1m   3.89692    2.81803  1.3829 1.701e-001
       totyrs  -0.08547    0.33847 -0.2525 8.012e-001
         ld72  -0.02859    0.07836 -0.3648 7.161e-001
         ld73  -0.18472    0.13462 -1.3721 1.734e-001
```

```
anova(f, ld72, ld73)   # get F with 2 numerator d.f.
```

The partial SSR for ld72,ld73 is now 462 instead of 748. $\frac{748-462}{748} = 0.38$ of the variation in maxfwt explained by ld72, ld73 has now been explained by the distance and years lived near the smelting plant.

6
```
anova(f, area, totyrs, ld72, ld73)
```

```
Factor     d.f. Partial SS        MS    F      P
 area        2    359.71835 179.85918 2.00 0.1417
 totyrs      1      5.74327   5.74327 0.06 0.8012
 ld72        1     11.98584  11.98584 0.13 0.7161
 ld73        1    169.57446 169.57446 1.88 0.1734
 REGRESSION  5   1111.88004 222.37601 2.47 0.0382  Answer here
 ERROR      91   8196.00708  90.06601
```

$H_0$: none of area, totyrs, ld72, ld73 are associated with maxfwt after adjusting for age, sex. $P = 0.04$ indicates some evidence for an association with at least one of the 4 exposure

variables. D.f. = 5 from area (2 from 3 groups), totyrs (1 since we assume linearity), ld72 (1, linear), ld73 (1, linear).