ST505: Fall 2012 Homework 6 solution

1. Consider the patient satisfaction data in problem 6.15 in the text.

   (a) Both $\mathbf{Y}$ and $\boldsymbol{\epsilon}$ are 43×1 vectors.

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 43 \\ 57 \\ 66 \\ . \\ . \\ 68 \\ 59 \\ 92 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ . \\ \epsilon_{46} \end{bmatrix}.$$

   (b) **MODEL 1:** $\mathbf{X}$ is 46×2, $\boldsymbol{\beta}$ and $\mathbf{b}$ are both 2×1 vectors and $\sigma^2\{\mathbf{b}\}$ is a 2×2 matrix.

$$\mathbf{X} = \begin{bmatrix} 1 & 50 \\ 1 & 36 \\ . & . \\ . & . \\ 1 & 37 \\ 1 & 28 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}, \quad \sigma^2\{\mathbf{b}\} = \begin{bmatrix} \sigma^2\{b_0\} & \sigma\{b_0, b_1\} \\ \sigma\{b_1, b_0\} & \sigma^2\{b_1\} \end{bmatrix}.$$

   MODEL 2: $\mathbf{X}$ is 46×4, $\boldsymbol{\beta}$ and $\mathbf{b}$ are both 4×1 vectors and $\sigma^2\{\mathbf{b}\}$ is a 4×4 matrix.

$$\mathbf{X} = \begin{bmatrix} 1 & 50 & 51 & 2.3 \\ 1 & 36 & 46 & 2.3 \\ . & . & & \\ . & . & & \\ 1 & 37 & 53 & 2.1 \\ 1 & 28 & 46 & 1.8 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix}, \quad \sigma^2\{\mathbf{b}\} = \begin{bmatrix} \sigma^2\{b_0\} & \sigma\{b_0, b_1\} & \sigma\{b_0, b_2\} & \sigma\{b_0, b_3\} \\ \sigma\{b_1, b_0\} & \sigma^2\{b_1\} & \sigma\{b_1, b_2\} & \sigma\{b_1, b_3\} \\ \sigma\{b_2, b_0\} & \sigma\{b_2, b_1\} & \sigma^2\{b_2\} & \sigma\{b_2, b_3\} \\ \sigma\{b_3, b_0\} & \sigma\{b_3, b_2\} & \sigma\{b_3, b_2\} & \sigma^2\{b_3\} \end{bmatrix}.$$

   MODEL 3: As with model $\mathbf{X}$ is 46×4, $\boldsymbol{\beta}$ and $\mathbf{b}$ are both 4×1 vectors and $\sigma^2\{\mathbf{b}\}$ is a 4×4 matrix. The form of $\sigma^2\{\mathbf{b}\}$ is exactly the same a model 2. All that changes here is what places the role of X_3 and so what is in the fourth column of $\mathbf{X}$.

$$\mathbf{X} = \begin{bmatrix} 1 & 50 & 51 & 50 * 51 \\ 1 & 36 & 46 & 36 * 46 \\ . & . & & \\ . & . & & \\ 1 & 37 & 53 & 37 * 53 \\ 1 & 28 & 46 & 28 * 46 \end{bmatrix}.$$

2. Fitting the Patient satisfaction data. Have shown R commands and output in most places below. Corresponding SAS code and output at end of solution.

   (a) `> fit3 <- lm(Satisfaction~Age +Severity + Anxiety)`

   The estimates, standard errors, confidence intervals and t-tests for the four coefficients are:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
Age          -1.1416     0.2148  -5.315 3.81e-06 ***
Severity     -0.4420     0.4920  -0.898   0.3741
Anxiety     -13.4702     7.0997  -1.897   0.0647 .
```

```
> confint(fit3)
                 2.5 %      97.5 %
(Intercept) 121.911727 195.0707761
Age          -1.575093  -0.7081303
Severity     -1.434831   0.5508228
Anxiety     -27.797859   0.8575324
```

The estimate of $\sigma^2\{\mathbf{b}\}$ assuming equal variances of error terms is $s^2\{\mathbf{b}\}$, given by:

```
> vcov(fit3)
            (Intercept)         Age    Severity    Anxiety
(Intercept) 328.5478428  0.93283693 -6.87207388 -6.8081417
Age           0.9328369  0.04613853 -0.03223004 -0.4716488
Severity     -6.8720739 -0.03223004  0.24203030 -1.7916031
Anxiety      -6.8081417 -0.47164876 -1.79160306 50.4051837
```

allowing the variance of the errors to change over observations the estimate is $s^2_{White}\{\mathbf{b}\} =$

```
> acov ## variance covariance matrix without assumption
            (Intercept)         Age    Severity    Anxiety
(Intercept) 277.7160961  0.99014977 -6.67977869  9.7237773
Age           0.9901498  0.04156309 -0.02728626 -0.5548634
Severity     -6.6797787 -0.02728626  0.23094574 -1.6780348
Anxiety       9.7237773 -0.55486345 -1.67803483 41.8845874
```

The analysis of variance table given by SAS is

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	9120.46367	3040.15456	30.05	<.0001
Error	42	4248.84068	101.16287		
Corrected Total	45	13369			

As noted, in class, the anova command in R (here anova(fit3)) does not give the anova table above. The F-statistic that corresponds to the anova comes from the summary command in R, leading to

```
> summary(fit3)
....
F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

The anova command in R gives

```
> anova(fit3)
Analysis of Variance Table
Response: Satisfaction
          Df Sum Sq Mean Sq F value    Pr(>F)
Age        1 8275.4  8275.4 81.8026 2.059e-11 ***
Severity   1  480.9   480.9  4.7539   0.03489 *
```

```
Anxiety    1  364.2   364.2  3.5997   0.06468 .
Residuals 42 4248.8   101.2
```

You could construct the traditional anova table (as given in SAS and most other software) using this via SSR = sum of the three one degree of freedom Sum SQ's; that is, SSR = 8275.5 +480.9 + 364.2 (subject to a little rounding difference). In fact, what the anova in R is giving you are additional sums of squares $SSR(X_1) = 8275.4$, $SSR(X_2|X_1) = 480.9$ and $SSR(X_3|X_1, X_2) = 364.2$.

(b) b_j is the estimate of the change in the expected value of Y when the jth predictor changes by 1 with the other two predictors held fixed. So, for example, $b_1 = .9328$ estimates the change in the expected satisfaction to be .9328 when age changes by 1 year while severity and age are held fixed.

(c) The tests associated with b_j is testing whether $\beta_j = 0$ in the model with $E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i2} + \beta_3 X_{i3}$.
At $\alpha = .05$ only the test for $H_0 : \beta_1 = 0$ leads to rejection, while at $\alpha = .10$ we reject for both β_1 and β_3. Since the p-value is only approximate (since the normality assumption is never exactly true), this points out the problem of working with a fixed α and having to make a yes(reject) or no (do not reject) decision.

(d) If the 95% confidence interval for β_j contains 0, then the p-value for the associated test will be greater than .05 (i.e., we would not reject $\beta_j = 0$). Conversely if 0 is NOT in the interval the the p-value will be less than .05 (i.e., we would reject $\beta_j = 0$).

(e) Interpret the F-test in the analysis of variance table. *As noted in class in R, we modified this question to be interpret the F-test from the summary command.* This is testing $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. The F-statistic is 30.05, based on 3 and 42 degrees of freedom, with a a P-value of $1.542e - 10$. This leads to rejecting H_0.

(f) All of the plots (residuals versus each of the three X's and fitted value and versus the three products, show systematic patterns indicating the linear regression model with the three X's appears to be a good fit and products are not needed.

Problems for ST697R students.

3. The fit with the three *original* terms and the three products yields

```
> fit5 <- lm(Satisfaction~Age +Severity + Anxiety + Age*Severity + Age*Anxiety + Severity*Anxiety)
> summary(fit5)
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      190.51810  117.37011   1.623    0.113
Age                0.79293    3.15488   0.251    0.803
Severity          -3.14572    3.26554  -0.963    0.341
Anxiety          -14.40686   70.96754  -0.203    0.840
Age:Severity       0.01565    0.06396   0.245    0.808
Age:Anxiety       -1.19694    0.93509  -1.280    0.208
Severity:Anxiety   0.93330    1.54466   0.604    0.549
```

The t-tests associated with the products test for the coefficients one-a-time, not simultaneous. None of these are rejected.

Not asked for You can test that $\beta_5 = \beta_6 = \beta_7 = 0$ (no interaction terms) using the general test using the test command in SAS or anova(fit3,fit5) in R after fitting the full and reduced model. Assuming equal variance this leads to F = 0.58 with 3 and 42 degrees of freedom and a p-value of .6339, so do not reject. The chi-square test from SAS allowing unequal variances reaches the same conclusion.

```
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     42 4248.8
2     39 4068.4  3    180.43 0.5765 0.6339

               Test noprod Results for Dependent Variable satis
                                         Mean
           Source             DF        Square    F Value    Pr > F
           Numerator           3      60.14187       0.58    0.6339
           Denominator        39     104.31833
                  Test noprod Results using Heteroscedasticity
                         Consistent Covariance Estimates
                           DF    Chi-Square    Pr > ChiSq
                            3          2.69        0.4421
```

4. Problem 6.22

   a) This is a multiple linear regression model as given (linear in β's although not in the X's)

   b) Taking natural log, leading to $Y_i^* = log(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + log(\epsilon_i)$

   This is not quite a linear model since with $E(\epsilon_i) = 0$, $E(log(\epsilon_i)) = \gamma \neq 0$. But, if we add in - γ we get $Y_i^* = log(Y_i) = \beta_0^* + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i^*$, where $\beta_0^* = \beta_0 - \gamma$ and ϵ_i^* had mean 0.

   c) Since $log_{10}(\beta_0 X_{i1}) = log_{10}(\beta_0) + log_{10}(X_{i1})$ you can write this as $Y_i - log_{10}(X_{i1}) = log_{10}(\beta_0) + \beta_2 X_{i2} + \epsilon_i$ or with the right definitions $Y_i^* = \beta_0^* + \beta_1^* X_{i2} + \epsilon_i$.

   d) is not "linearizable".

   e) $Y_i^* = log((1/Y_i) - 1) = \beta_0 + \beta_1 X_{i1} + \epsilon_i$, which is a linear model.

   NOTE: If we fit the transformed models in b) or e) and estimate the mean we are estimating $E(Y^*)$. If we want to estimate $E(Y)$ it is not as simple as just transforming back since the transformation is non-linear. For example, in b) suppose $\widehat{\mu}^*$ estimates $E(Y^*)$ at some set of $X's$. Then $e^{\widehat{\mu}^*}$ is NOT unbiased for $E(Y)$ and we can't just transform the intervals for μ^*. The problem is that exponentiation is a non-linear function. There are ways to use approximates to estimate $E(Y)$. *However if the problem is just prediction, then we can do a prediction interval for Y^* and transform back to get a prediction interval for Y.*

5. Problem 6.16 b). Hint: Use Bonferroni's method.

   Use $b_j \pm t(1 - (.10/6), 42)s\{b_j\}$. This leads to the following simultaneous intervals:

```
beta1  -1.614248 -0.6689755
beta2  -1.524510  0.6405013
beta3 -29.092028  2.1517012
```

```
SAS code and output.

title 'patient example, prob 6.15 in NWNK ';
options  linesize=80;
data a;
infile 'g:/s505/data/patient5.txt';
```

```
input satis age severity anxiety;
x1x2=age*severity;
x1x3 = age*anxiety;
x2x3 = severity*anxiety;
run;
proc reg;
model satis = age severity anxiety/covb acov clb;
run;
title 'simultaneous CIs for beta1, beta2 and beta3';
proc reg; /* this will automatically give simulteneous
           90% CI's for the three non-intercept coefficients */
model satis = age severity anxiety/clb alpha = .0333333;
run;
proc reg;
model satis = age severity anxiety x1x2 x1x3 x2x3/covb acov;
noprod: test x1x2=0, x1x3=0, x2x3=0;
run;
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	9120.46367	3040.15456	30.05	<.0001
Error	42	4248.84068	101.16287		
Corrected Total	45	13369			

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	158.49125	18.12589	8.74	<.0001
age	1	-1.14161	0.21480	-5.31	<.0001
severity	1	-0.44200	0.49197	-0.90	0.3741
anxiety	1	-13.47016	7.09966	-1.90	0.0647

--Heteroscedasticity Consistent-

Variable	DF	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	16.66482	9.51	<.0001	121.91173	195.07078
age	1	0.20387	-5.60	<.0001	-1.57509	-0.70813
severity	1	0.48057	-0.92	0.3630	-1.43483	0.55082
anxiety	1	6.47183	-2.08	0.0435	-27.79786	0.85753

Heteroscedasticity Consistent

Variable	DF	95% Confidence Limits	
Intercept	1	124.86029	192.12221
age	1	-1.55304	-0.73018
severity	1	-1.41183	0.52782
anxiety	1	-26.53085	-0.40948

Covariance of Estimates

Variable	Intercept	age	severity	anxiety
Intercept	328.54784276	0.9328369266	-6.872073881	-6.808141658
age	0.9328369266	0.0461385284	-0.032230039	-0.471648757
severity	-6.872073881	-0.032230039	0.2420302972	-1.791603061
anxiety	-6.808141658	-0.471648757	-1.791603061	50.405183679

Heteroscedasticity Consistent Covariance of Estimates

Variable	Intercept	age	severity	anxiety
Intercept	277.71609611	0.99014977	-6.679778695	9.7237772911
age	0.99014977	0.0415630924	-0.027286261	-0.554863448
severity	-6.679778695	-0.027286261	0.2309457396	-1.678034827
anxiety	9.7237772911	-0.554863448	-1.678034827	41.884587402

simultaneous CIs for beta1, beta2 and beta3

		96.66667% Confidence	
Variable	DF	Limits	
Intercept	1	118.60762	198.37488
age	1	-1.61425	-0.66898
severity	1	-1.52451	0.64050
anxiety	1	-29.09203	2.15170

Fitting with products

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	190.51810	117.37011	1.62	0.1126
age	1	0.79293	3.15488	0.25	0.8029
severity	1	-3.14572	3.26554	-0.96	0.3413
anxiety	1	-14.40686	70.96754	-0.20	0.8402
x1x2	1	0.01565	0.06396	0.24	0.8080
x1x3	1	-1.19694	0.93509	-1.28	0.2081
x2x3	1	0.93330	1.54466	0.60	0.5492

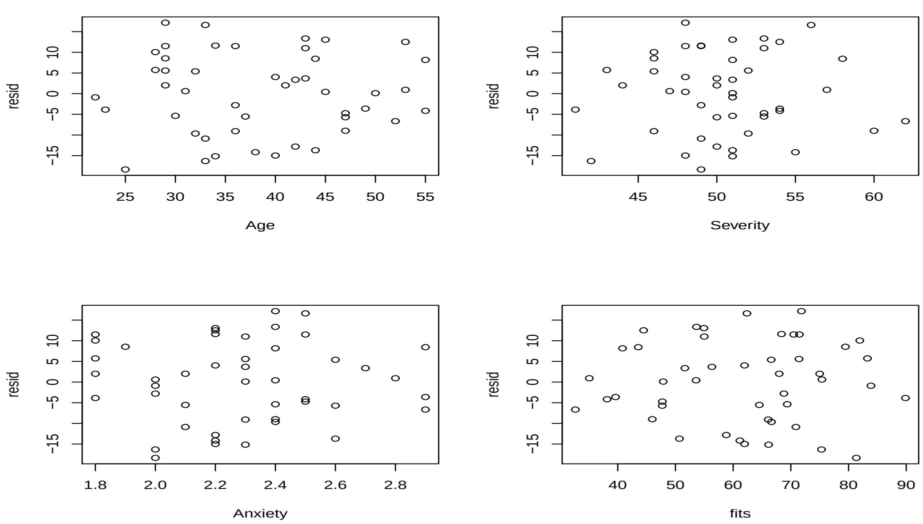

Figure 1: Residual plots for patient data.

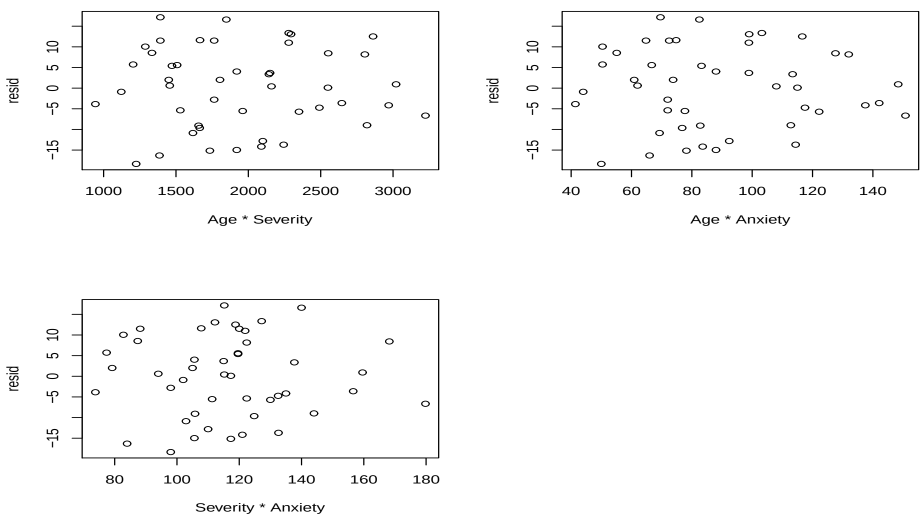

Figure 2: Residual plots for patient data; versus products.