**Name** **Fall, 2013**

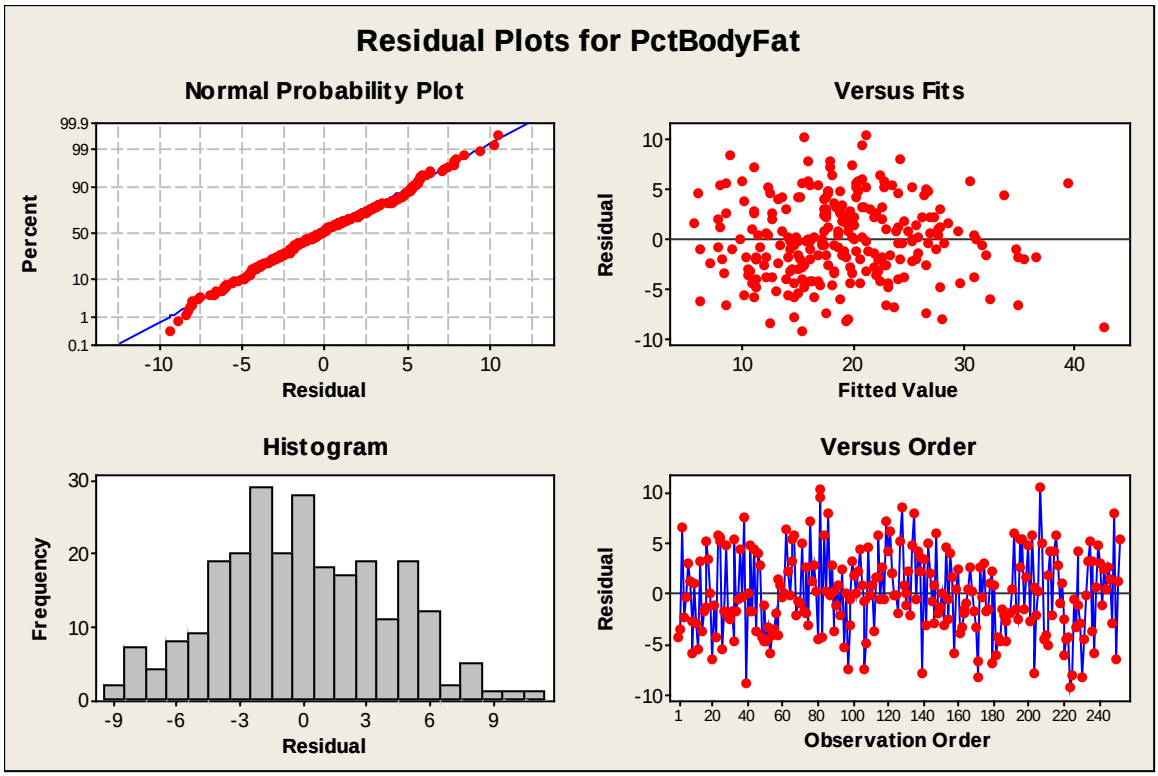
## Applied Statistics Comprehensive Examination
## Regression & Linear Models

1. (50 Points) A nutritionist is trying to estimate percent body fat based on the circumference of various body parts. A multiple regression of percent body fat on all 10 body part measures of circumference (neck, chest, abdomen, hip, thigh, knee, ankle, extended bicep, forearm, and wrist) was performed and the following output was produced.

Analysis of Variance

| Source | DF | SS |
|---|---|---|
| Regression | 10 | 11084.3 |
| Residual Error | 241 | 3994.7 |
| Total | 251 | 15079.0 |

| Predictor | Coef | SE Coef | T | P | VIF |
|---|---|---|---|---|---|
| Constant | 7.229 | 6.214 | 1.16 | 0.246 | |
| Neck Circumference | -0.5819 | 0.2086 | -2.79 | 0.006 | 3.893 |
| Chest Circumference | -0.09085 | 0.08543 | -1.06 | 0.289 | 7.855 |
| Abdomen Circumference | 0.96023 | 0.07158 | 13.41 | 0.000 | 9.022 |
| Hip Circumference | -0.3914 | 0.1127 | -3.47 | 0.001 | 9.869 |
| Thigh Circumference | 0.1337 | 0.1249 | 1.07 | 0.286 | 6.513 |
| Knee Circumference | -0.0941 | 0.2124 | -0.44 | 0.658 | 3.973 |
| Ankle Circumference | 0.0042 | 0.2032 | 0.02 | 0.983 | 1.796 |
| Extended Biceps Circumference | 0.1112 | 0.1591 | 0.70 | 0.485 | 3.500 |
| Forearm Circumference | 0.3445 | 0.1855 | 1.86 | 0.064 | 2.128 |
| Wrist Circumference | -1.3535 | 0.4714 | -2.87 | 0.004 | 2.933 |


Residual Plots for PctBodyFat
Normal Probability Plot
Percent
99.9
99
90
50
10
1
0.1
-10
-5
0
5
10
Residual
Versus Fits
Residual
10
5
0
-5
-10
10
20
30
40
Fitted Value
Histogram
Frequency
30
20
10
0
-9
-6
-3
0
3
6
9
Residual
Versus Order
Residual
10
5
0
-5
-10
1
20
40
60
80
100
120
140
160
180
200
220
240
Observation Order

a. (5 points) Calculate and interpret the value of the $R^2$ for this model.

b. (15 points) State all the assumptions for this multiple regression. For each one, provide evidence of whether the assumption is met. If you cannot verify it from the given output, state what additional output you would need in order to verify it.

c. (10 points) The following is the output from another model that includes only a subset of the variables. What hypothesis test is being tested by a partial F comparing this model to the one above? Conduct the test at the 5% level of significance.

| Predictor | Coef | SE Coef | T | P |
|---|---|---|---|---|
| Constant | 2.167 | 5.776 | 0.38 | 0.708 |
| Neck Circumference | -0.7146 | 0.1829 | -3.91 | 0.000 |
| Chest Circumference | -0.05929 | 0.08349 | -0.71 | 0.478 |
| Abdomen Circumference | 0.93565 | 0.07132 | 13.12 | 0.000 |
| Hip Circumference | -0.4185 | 0.1142 | -3.66 | 0.000 |
| Thigh Circumference | 0.2593 | 0.1175 | 2.21 | 0.028 |
| Knee Circumference | -0.1929 | 0.2092 | -0.92 | 0.358 |
| Ankle Circumference | -0.1237 | 0.1997 | -0.62 | 0.536 |

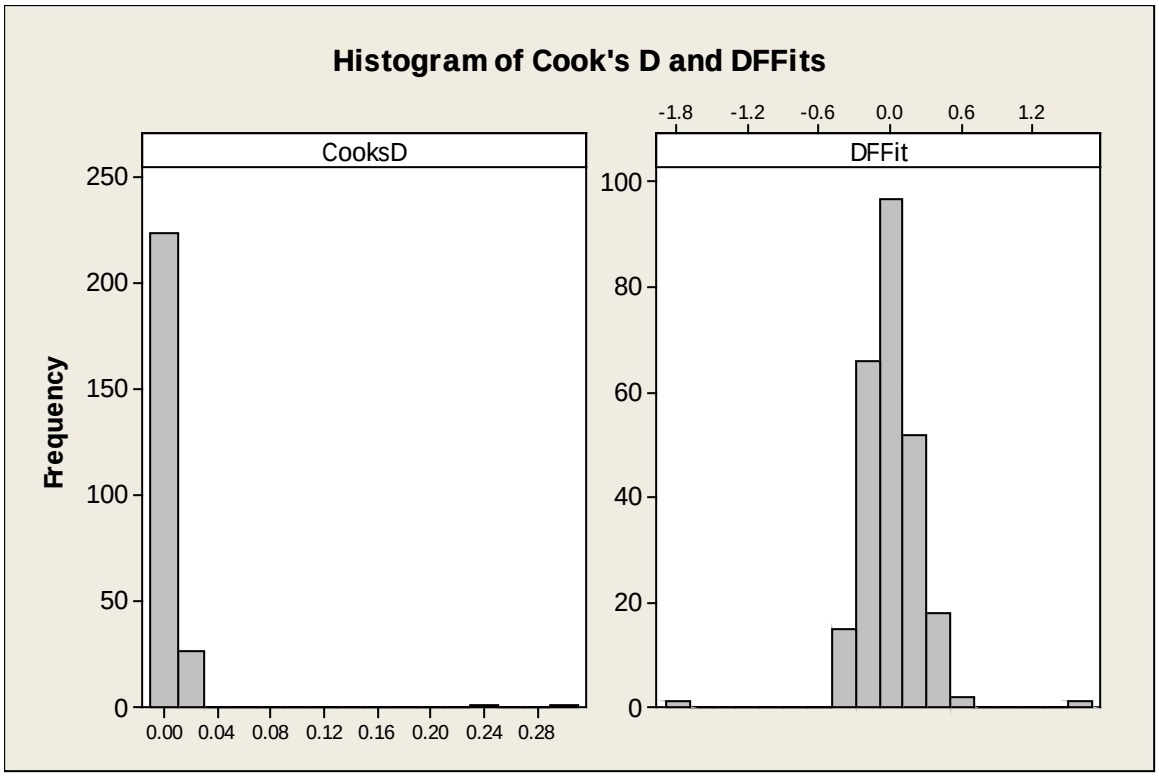
Analysis of Variance

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 7 | 10896.1 | 1556.6 | 90.80 | 0.000 |
| Residual Error | 244 | 4182.9 | 17.1 | | |
| Total | 251 | 15079.0 | | | |

d. (10 points) Consider only the abdomen, wrist, and forearm circumference variables. Based on the following output, conduct the stepwise procedure using both $\alpha_{entry}=0.05$ and $\alpha_{stay}=0.05$ to determine which, if any, of these three variables would be selected in a final model.

```
Predictor                Coef  SE Coef       T      P
Constant               -35.197    2.462  -14.29  0.000
Abdomen Circumference  0.58489  0.02643   22.13  0.000

Predictor               Coef  SE Coef      T      P
Constant             -33.666    8.987  -3.75  0.000
Wrist Circumference   2.8856   0.4923   5.86  0.000

Predictor                Coef  SE Coef      T      P
Constant               -21.003    6.495  -3.23  0.001
Forearm Circumference   1.3934   0.2260   6.17  0.000

Predictor                Coef  SE Coef      T      P
Constant                -7.162    5.379  -1.33  0.184
Abdomen Circumference  0.69833  0.03169  22.04  0.000
Wrist Circumference    -2.1138   0.3660  -5.78  0.000

Predictor                Coef  SE Coef      T      P
Constant               -30.458    4.071  -7.48  0.000
Abdomen Circumference  0.60731  0.03051  19.90  0.000
Forearm Circumference  -0.2377   0.1628  -1.46  0.146

Predictor                Coef  SE Coef      T      P
Constant               -38.849    8.936  -4.35  0.000
Wrist Circumference     1.7037   0.5950   2.86  0.005
Forearm Circumference   0.9325   0.2749   3.39  0.001

Predictor                Coef  SE Coef      T      P
Constant                -8.192    5.515  -1.49  0.139
Abdomen Circumference  0.69220  0.03250  21.30  0.000
Wrist Circumference    -2.2518   0.4002  -5.63  0.000
Forearm Circumference   0.1435   0.1679   0.85  0.394
```

e. (10 points) What do the following plots of Cook's Distance and DFFits tell you about the data? These are based on the full model with all 10 independent variables.



2. (20 Points) The data below were obtained from a study of the effect of oven temperature and baking time on the life (in hours) of an electrical component.

| Oven Temperature (F) | Baking Time (min) 5 | 10 | 15 | Mean |
|---|---|---|---|---|
| 600 | 165 | 192 | 170 | 201.33 |
| | 220 | 224 | 203 | |
| | 212 | 228 | 198 | |
| 620 | 170 | 172 | 168 | 188.89 |
| | 192 | 220 | 189 | |
| | 181 | 213 | 195 | |
| 640 | 123 | 132 | 116 | 141.00 |
| | 160 | 134 | 143 | |
| | 149 | 153 | 159 | |
| Mean | 174.67 | 185.33 | 171.22 | |

Minitab output for the two-way design with interaction is given below.

**Two-way ANOVA: hours versus temp, time**

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| temp | 2 | 18265.0 | 9132.48 | 23.15 | 0.000 |
| time | 2 | 974.3 | 487.15 | 1.23 | 0.314 |
| Interaction | 4 | 727.3 | 181.81 | 0.46 | 0.763 |
| Error | 18 | 7101.3 | 394.52 | | |
| Total | 26 | 27067.9 | | | |

a. (10 Points) Using only one observation per cell, write the design matrix for the effects model using sum-to-zero restrictions.
b. (10 Points) Give a complete set of orthogonal contrasts for oven temperature and explain what each contrast is testing. Carry out the each of the tests at the 0.05 level of significance and state your conclusions.

3. (30 Points) A small greenhouse experiment was carried out to investigate the effects of two soil types on the yield of two varieties of tomato. The data are shown below.

| Soil Type | Variety A | Variety B |
|---|---|---|
| Type 1 | 50 | 59 |
| | 55 | 56 |
| | 48 | 53 |
| | 53 | |
| Type 2 | 54 | 62 |
| | 52 | 58 |
| | | 59 |

a. (10 Points) Calculate the means and least squares means for each variety of tomato.

b. (10 Points) Consider the effects model: $y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk,}$ where $\alpha_i$ is the effect of the ith soil type, $\beta_j$ is the effect of the jth variety, and $\gamma_{ij}$ is the interaction between the ith soil type and jth variety. Determine if the following are estimable. Explain why or why not:

   i. $\beta_1 - \beta_2$

   ii. $\gamma_{11} - \gamma_{12} - \gamma_{21} + \gamma_{22}$

c. (10 Points) The above data had several observations deleted for the purpose of answering questions 3a and 3b. Suppose the data from the full experiment are found in the following table and that the two soil types were randomly selected for the experiment.

| Soil Type | Variety A | Variety B |
|---|---|---|
| Type 1 | 50 | 59 |
| | 55 | 56 |
| | 48 | 53 |
| | 53 | 58 |
| Type 2 | 54 | 62 |
| | 52 | 58 |
| | 50 | 59 |
| | 57 | 53 |

State the sources of variation and the associated degrees of freedom and expected mean square for each term. Briefly describe how to test the significance of variety. Do not actually perform the test.