CSC 423/324 – Early Final Exam

February 22, 2012

Part A. Multiple Choice Problems. 3 pts. each. Answer 19 of 20 questions. For each question give a reason or show your work for possible partial credit. ***For starred problems a reason or work is required***.

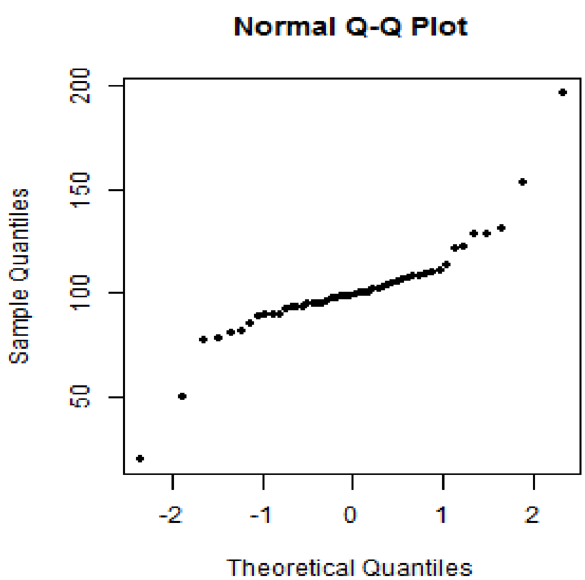
1. What is the definition of a random variable?
   a. A process of choosing a random number.
   b. The extreme values of a normal distribution.
   c. The extreme values of a normally distributed dataset.
   d. The median of a normal density.

2. * The horizontal distance between the inflection points of a normal density for a population is
   a. σ b. 2σ c. σ^2 d. $\mu + \sigma$

3. * What is the interquartile range for a dataset for a population that has an exactly standard normal distribution?
   a. 0.67 b. 1.00 c. 1.35 d. 1.96

4. For a continuous symmetric probability density, denote the population mean by μ, the population median by ν, the sample mean by x, and the sample median by m. Which of the following is true?
   a. $\mu = \nu$ b. $\mu < \nu$ c. $\mu > \nu$ d. x = m

5. What does the normal plot below tell you about the probability histogram of the dataset?
   a. It is skewed to the left. b. It is skewed to the right.
   c. It has thin tails d. It has thick tails.



6. What is the most important reason why data from observational studies more difficult to interpret than data obtained from an experiment with treatments randomly assigned to subjects? It is hard to
   a. create a normal plot with data from observational studies.
   b. distinguish dependent variables from independent variables.
   c. find subjects willing to participate in a randomized study.
   d. tell if the effect being studied is due to the treatments or is due to some variable not included as an independent variable in the model.

7. * A dataset having 16 observations is normally distributed with $\bar{x}$ = 4.52 and s_x = 0.345. Find a 99% confidence interval for the true value of μ
   a. (3.92, 5.12)   b. (4.26, 4.78)   c. (4.37, 4.67)   d. (4.26, 4.77)

8. Which of these is the normal equation for a regression model in matrix form?
   a. $\mathbf{X}^{-1}\,\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^{-1}\,\mathbf{y}$ b. $\mathbf{X}^{T}\,\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^{T}\,\mathbf{y}$ c. $\mathbf{X}^{-1}\,\mathbf{X}\beta = \mathbf{X}\text{-}1\,\mathbf{y}$ d. $\mathbf{X}\sigma^2 = \boldsymbol{\mu}$

9. * A random sample of 15 college age women agrees to take fish oil for one year, and then take an IQ test. Sample mean for the IQ scores is 115 with s_x = 12. The average IQ score at the college that the women attend is 108. Do you accept the null hypothesis that taking the fish oil did not make a difference in intelligence at the 0.10 level? At the 0.05 level?
   a. no; no   b. no; yes   c. yes; no   d. yes; yes

10. * Which of the following could be most easily tested with a paired two -sample t-test? Give a reason to support your answer.
    a. Whether automobile Model A gets better gas mileage than automobile Model B.
    b. Whether Candidate A or Candidate B is more likely to win the election next month.
    c. Whether a new drug lowers chlorestoral better than a currently popular drug.
    d. Which of two websites is easier to use.

11. An estimated regression parameter is unbiased if
    a. its expected value is equal to the value of the corresponding true regression parameter.
    b. its standard error is smaller than any other estimated regression parameter.
    c. only if the regression model is regression through the origin.
    d. the MSE for the regression model equals $(n\text{-}1)s_x^2$

12. * What is the 5th normal score computed using Van der Waerden's method if n = 9?
    a. – 0.253   b. 0.000   c. 0.140   d. 0.500

13. * For a simple linear regression model, if s_x = 25.3, s_y = 31.7, and s_{xy} = 361, the R-squared value is
    a. 0.20   b. 0.27   c. 0.33   d. 0.45

Use the following SAS output to answer Questions 13, 14, and 15.

The i option in the SAS model statement produces these values for $(\mathbf{X}^T\mathbf{X})^{-1}$:

```
X'X Inverse
Variable   Intercept       x1       x2
Intercept   0.89583   -0.1875  -0.0556
x1         -0.18750    0.0625   0.0000
x2         -0.05556    0.0000   0.0185
```

The xpx option in the SAS model statement gives these values for X'Y:

```
Model Crossproducts X'Y
Variable          y
Intercept      87.2
x1            274.0
x2            348.6
```

14. * What is the estimated intercept for the regression equation?
    a. 0.896 b. 7.36 c. 12.9 d. 87.2

15. * What is the estimated regression parameter associated with x2?
    a. 0.0185 b. 1.60 c. 3.48 d. 348.6

16. * If n = 6 and SSE = 0.7367, what is the standard error of the estimated regression parameter assodiated with x1?
    a. 0.625 b. 0.124 c. 0.246 d. 0.496

17. * For the regression equation at the top of this page, SSE=0.7367 and SST=149.78, what is the R-squared value?
    a. 0.9643 b. 0.9951 c. 0.9975 d. 1.0000

18. * For a regression model with 7 regression parameters (including the intercept) and n = 19, the value of the F statistic for the overall F test is 4.17. Are any regressors significant at the 0.05 level? At the 0.01 level?
    a. no; no b. yes; no c. no; yes d. yes; yes

19. If H is the hat matrix, then $(\mathbf{I} - \mathbf{H})$y represents the
    a. vector of predicted values.
    b. vector of residuals.
    c. standard errors of the estimated parameters.
    d. vector of estimated parameters.

20. A large variance inflation factor for an estimated parameter indicates
    a. Heteroscedasticity b. Multicollinarity
    c. A leverage point d. An outlier

Part B: Short Essay Questions. 10 pts. each. For full credit write in compete sentences and paragraphs. Do only 2 out of 3 questions.

1. Explain what the Central Limit Theorem is and why it is important for statistical tests.

2. What are influence or leverage points. How do they differ from outliers? How is information about influence points used to find a good regression model?

3. Assume a multiple regression model. Explain the difference between a confidence interval for $\hat{y}$ and a prediction interval for a new observation.

Part C: Short Answer Questions. Answer all questions about the Clinical Depression Dataset, using the output and plots on pages 5 to 16. Pages 5 to 8: SAS Output, pages 9 to 12: R Output, pages 13 to 16: Residual and Normal Plots. The variables in the dataset are age (Age of Patient), sex (Sex of Patient, 0=male, 1=female), wp (Work Place Conflict), mc (Marital Conflict), dep (Depression Score on Psychological Evaluation).

1. (5 pts.) Look at the plots on Page 13. Are there any outliers in either the group of males or the group of females? Explain your answer.

2. (10 pts.) Look at the SAS or R Output. Write out in detail the five steps of the independent 2-sample t-test for testing whether there is a difference in clinical depression rates for men vs. women at the 0.05 level. Compute a 95% confidence interval for the test statistic by hand, but obtain any other values from the SAS on Page 5 or R output on Page 9.

3. (5 pts.) What are the assumptions for the t-test in Question2? Do these assumptions appear to be met? Refer to the plots shown on Page 13?

4. (5 pts.) What are the values of the overall F-statistic and associated p-value for Model 1. Interpret them. What do they tell you about Model 1?

5. (15 pts.) Based on the SAS or R Output and the Diagnostic Plots, which of models 1 through 7 is the best regression model. Explain your answer. Include a comparison of the R-squared and adjusted R-squared values in your discussion. Here are the seven regression models:

   Model 1: dep=age wp mc   Model 2 : dep=wp mc   Model 3: dep=age mc   Model 4: age wp
   Model 5: dep=age   Model 6 : dep=wp   Model 7: dep=mc

6. (10 pts.) Models 1 through 4 contain information about multicollinearity. (a) What is this information and what does it tell you about the regression problems? (b) What should be done with a regression that has high multicollinearity for one or more variables? Why is information about multicollinearity not approriate for models 5, 6, and 7?

The TTEST Procedure

Variable: dep

sex	N	Mean	Std Dev	Std Err	Minimum	Maximum
0	16	157.6	68.8349	17.2087	37.0000	294.0
1	23	112.0	51.2174	10.6796	33.0000	238.0
Diff (1-2)		45.6685	58.9972	19.2061		

sex	Method	Mean	95% CL Mean		Std Dev
0		157.6	120.9	194.3	68.8349
1		112.0	89.8084	134.1	51.2174
Diff (1-2)	Pooled	45.6685	6.7532	84.5838	58.9972
Diff (1-2)	Satterthwaite	45.6685	4.0479	87.2891	

sex	Method	95% CL Std Dev	
0		50.8487	106.5
1		39.6113	72.4907
Diff (1-2)	Pooled	48.0983	76.3274
Diff (1-2)	Satterthwaite		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	37	2.38	0.0227
Satterthwaite	Unequal	26.136	2.25	0.0327

Equality of Variances

Method	Num DF	Den DF	F Value	Pr > F
Folded F	15	22	1.81	0.2020

The REG Procedure   Analysis of Variance   MODEL1: dep=age wp mc

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	53407	17802	6.55	0.0012
Error	35	95057	2715.92465		
Corrected Total	38	148464			

Root MSE	52.11453	R-Square	0.3597
Dependent Mean	130.69231	Adj R-Sq	0.3048
Coeff Var	39.87575		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	194.54366	91.76355	2.12	0.0412	0
age	1	-1.87579	0.89667	-2.09	0.0438	1.01513
wp	1	0.50993	1.16200	0.44	0.6635	1.01155
mc	1	-1.22887	0.30440	-4.04	0.0003	1.00485

The REG Procedure Analysis of Variance MODEL 2: dep=wp mc

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	41521	20761	6.99	0.0027
Error	36	106943	2970.64104		
Corrected Total	38	148464			

Root MSE	54.50359	R-Square	0.2797
Dependent Mean	130.69231	Adj R-Sq	0.2397
Coeff Var	41.70375		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	146.75412	92.94876	1.58	0.1231	0
wp	1	0.25565	1.20860	0.21	0.8337	1.00048
mc	1	-1.18690	0.31766	-3.74	0.0006	1.00048

The REG Procedure Analysis of Variance MODEL 3: dep=age mc

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	52884	26442	9.96	0.0004
Error	36	95580	2655.01074		
Corrected Total	38	148464			

Root MSE	51.52680	R-Square	0.3562
Dependent Mean	130.69231	Adj R-Sq	0.3204
Coeff Var	39.42604		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	231.90616	33.84431	6.85	<.0001	0
age	1	-1.83463	0.88169	-2.08	0.0446	1.00402
mc	1	-1.22503	0.30084	-4.07	0.0002	1.00402

The REG Procedure   Analysis of Variance   MODEL 4: dep=age wp

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	9143.98087	4571.99044	1.18	0.3185
Error	36	139320	3870.00908		
Corrected Total	38	148464			

Root MSE	62.20940	R-Square	0.0616
Dependent Mean	130.69231	Adj R-Sq	0.0095
Coeff Var	47.59989		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	159.61006	109.05051	1.46	0.1520	0
age	1	-1.63723	1.06803	-1.53	0.1340	1.01072
wp	1	0.37516	1.38651	0.27	0.7883	1.01072

The REG Procedure   Analysis of Variance   MODEL 5: dep=age

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8860.64696	8860.64696	2.35	0.1339
Error	37	139604	3773.07191		
Corrected Total	38	148464			

Root MSE	61.42534	R-Square	0.0597
Dependent Mean	130.69231	Adj R-Sq	0.0343
Coeff Var	46.99996		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	187.20122	38.16424	4.91	<.0001	0
age	1	-1.60747	1.04896	-1.53	0.1339	1.00000

The REG Procedure   Analysis of Variance   MODEL 6: dep=wp

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	49.70052	49.70052	0.01	0.9120
Error	37	148415	4011.20560		
Corrected Total	38	148464			

Root MSE	63.33408	R-Square	0.0003
Dependent Mean	130.69231	Adj R-Sq	-0.0267
Coeff Var	48.46045		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	118.76210	107.65660	1.10	0.2771	0
wp	1	0.15629	1.40408	0.11	0.9120	1.00000

The REG Procedure   Analysis of Variance   MODEL 7: dep=mc

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	41388	41388	14.30	0.0006
Error	37	107076	2893.94584		
Corrected Total	38	148464			

Root MSE	53.79541	R-Square	0.2788
Dependent Mean	130.69231	Adj R-Sq	0.2593
Coeff Var	41.16188		

Parameter Estimates

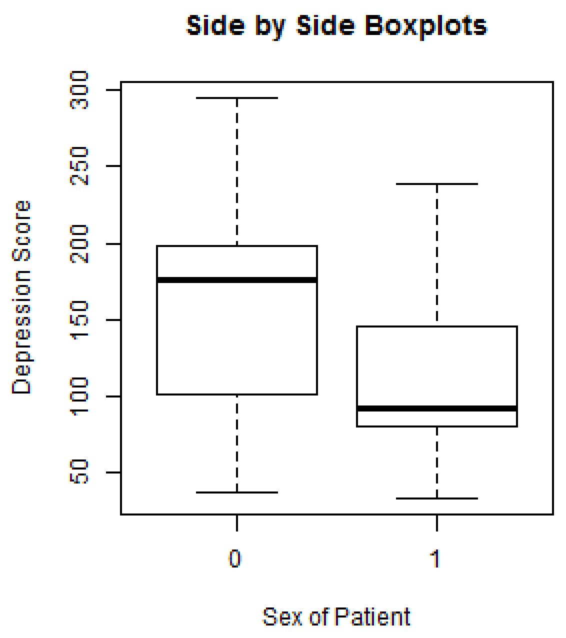
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	166.22468	12.74690	13.04	<.0001	
mc	1	-1.18543	0.31346	-3.78	0.0006	1.00000

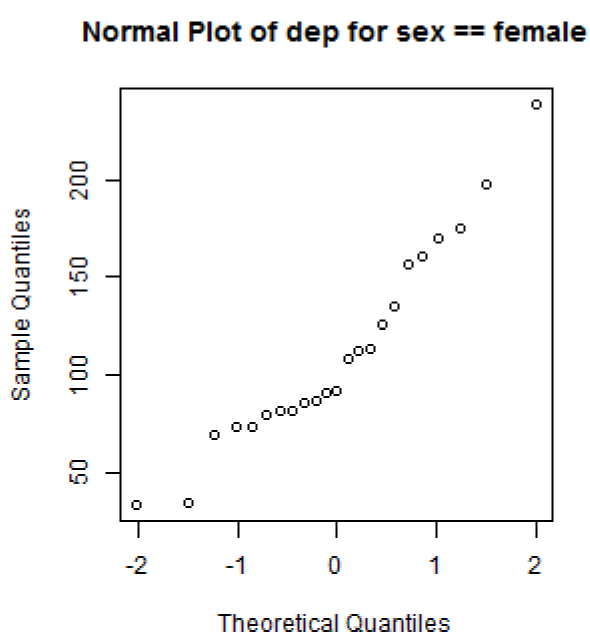
```
        Two Sample t-test

data:  dep by sex
t = 2.3778, df = 37, p-value = 0.02269
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  6.753179 84.583777
sample estimates:
mean in group 0 mean in group 1
       157.6250        111.9565

> t.test(dep ~ sex, var.equal=FALSE)

        Welch Two Sample t-test

data:  dep by sex
t = 2.2549, df = 26.136, p-value = 0.03274
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  4.047857 87.289099
sample estimates:
mean in group 0 mean in group 1
       157.6250        111.9565

> male = dep[sex==0]
> female = dep[sex==1]
> var.test(female, male)

        F test to compare two variances

data:  female and male
F = 0.5536, num df = 22, denom df = 15, p-value = 0.202
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.203095 1.383186
sample estimates:
ratio of variances
         0.5536275

-------------------------------------------------------------------------------------
```

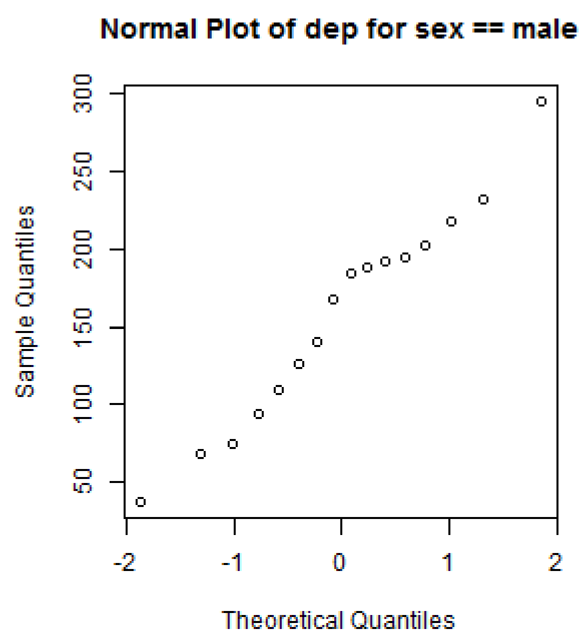
```
Summary for Model 1:

Call:
lm(formula = dep ~ age + wp + mc)

Residuals:
    Min      1Q  Median      3Q     Max
-95.889 -29.683  -5.643  37.246 134.027

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 194.5437    91.7635   2.120  0.04116 *
age          -1.8758     0.8967  -2.092  0.04376 *
wp            0.5099     1.1620   0.439  0.66348
mc           -1.2289     0.3044  -4.037  0.00028 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52.11 on 35 degrees of freedom
Multiple R-squared: 0.3597,     Adjusted R-squared: 0.3048
F-statistic: 6.555 on 3 and 35 DF,  p-value: 0.001236

Variance inflation factors for Model 1:
     age       wp       mc
1.015126 1.011552 1.004849

-------------------------------------------------------------------------------------

Summary for Model 2:

Call:
lm(formula = dep ~ wp + mc)

Residuals:
    Min      1Q  Median      3Q     Max
-91.669 -44.620   1.125  30.189 134.919

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 146.7541    92.9488   1.579 0.123113
wp            0.2557     1.2086   0.212 0.833668
mc           -1.1869     0.3177  -3.736 0.000646 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 54.5 on 36 degrees of freedom
Multiple R-squared: 0.2797,     Adjusted R-squared: 0.2397
F-statistic: 6.989 on 2 and 36 DF,  p-value: 0.002726

Variance inflation factors for Model 2:
      wp       mc
1.000484 1.000484

-------------------------------------------------------------------------------------
```

```
Summary for Model 3:

Call:
lm(formula = dep ~ age + mc)

Residuals:
    Min      1Q  Median      3Q     Max 
-98.997 -31.791  -6.979  35.465 141.000 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 231.9062    33.8443   6.852 5.14e-08 ***
age          -1.8346     0.8817  -2.081 0.044626 *  
mc           -1.2250     0.3008  -4.072 0.000244 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51.53 on 36 degrees of freedom
Multiple R-squared: 0.3562,     Adjusted R-squared: 0.3204 
F-statistic: 9.959 on 2 and 36 DF,  p-value: 0.0003609

Variance inflation factors for Model 3:
     age       mc 
1.004019 1.004019 

-------------------------------------------------------------------------------------


Summary for Model 4:

Call:
lm(formula = dep ~ age + wp)

Residuals:
    Min      1Q  Median      3Q     Max 
-119.85  -43.02  -12.39   46.98  161.20 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 159.6101   109.0505   1.464    0.152
age          -1.6372     1.0680  -1.533    0.134
wp            0.3752     1.3865   0.271    0.788

Residual standard error: 62.21 on 36 degrees of freedom
Multiple R-squared: 0.06159,    Adjusted R-squared: 0.009457 
F-statistic: 1.181 on 2 and 36 DF,  p-value: 0.3185

Variance inflation factors for Model 4:
     age       wp 
1.010718 1.010718 

-------------------------------------------------------------------------------------
```

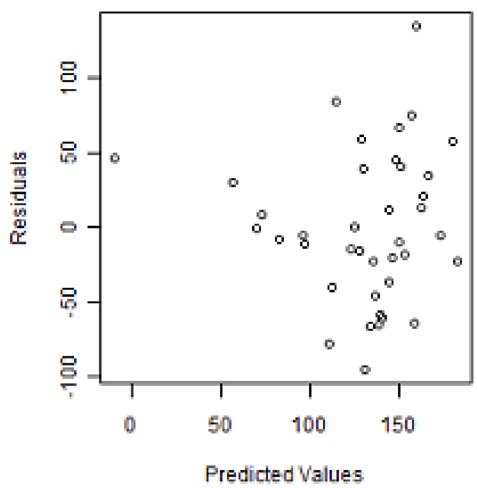
```
Summary for Model 5:
Call:
lm(formula = dep ~ age)

Residuals:
    Min      1Q  Median      3Q     Max
-118.84  -42.99  -11.37   46.79  166.28

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  187.201     38.164   4.905 1.89e-05 ***
age           -1.607      1.049  -1.532    0.134
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61.43 on 37 degrees of freedom
Multiple R-squared: 0.05968,    Adjusted R-squared: 0.03427
F-statistic: 2.348 on 1 and 37 DF,  p-value: 0.1339

-------------------------------------------------------------------------------------
Summary for Model 6:
Call:
lm(formula = dep ~ wp)

Residuals:
   Min     1Q Median     3Q    Max
-97.95 -48.48 -17.48  49.02 161.17

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 118.7621   107.6566   1.103    0.277
wp            0.1563     1.4041   0.111    0.912

Residual standard error: 63.33 on 37 degrees of freedom
Multiple R-squared: 0.0003348,  Adjusted R-squared: -0.02668
F-statistic: 0.01239 on 1 and 37 DF,  p-value: 0.912

-------------------------------------------------------------------------------------
Summary for Model 7:
Call:
lm(formula = dep ~ mc)

Residuals:
    Min      1Q  Median      3Q     Max
-93.291 -46.334   0.855  29.871 138.444

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 166.2247    12.7469  13.040 2.09e-15 ***
mc           -1.1854     0.3135  -3.782 0.000551 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 53.8 on 37 degrees of freedom
Multiple R-squared: 0.2788,     Adjusted R-squared: 0.2593
F-statistic:  14.3 on 1 and 37 DF,  p-value: 0.0005512
-------------------------------------------------------------------------------------
```


Side by Side Boxplots
Depression Score
50
100
150
200
250
300
0
1
Sex of Patient

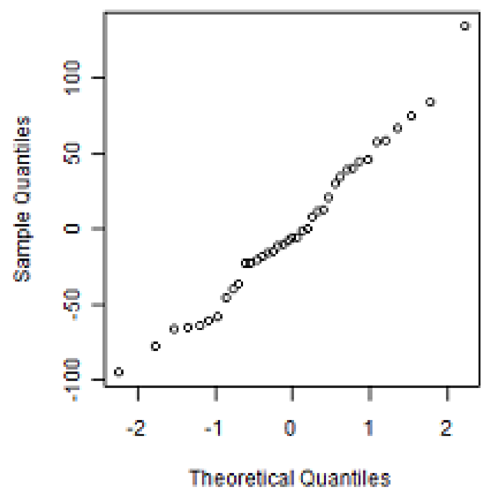

Normal Plot of dep for sex == female
Sample Quantiles
50
100
150
200
-2
-1
0
1
2
Theoretical Quantiles


Normal Plot of dep for sex == male
Sample Quantiles
50
100
150
200
250
300
-2
-1
0
1
2
Theoretical Quantiles

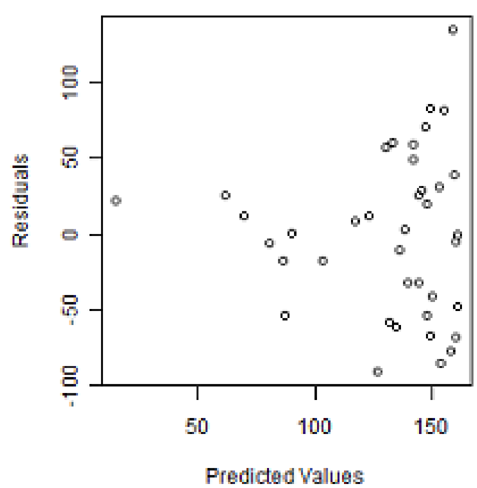
Model 1: Residual Plot



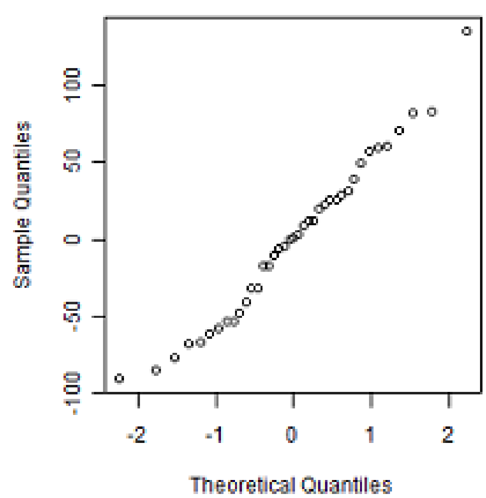
Model 1: Normal Plot



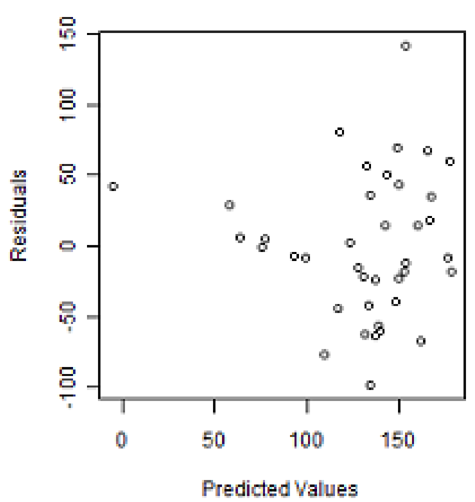
Model 2: Residual Plot



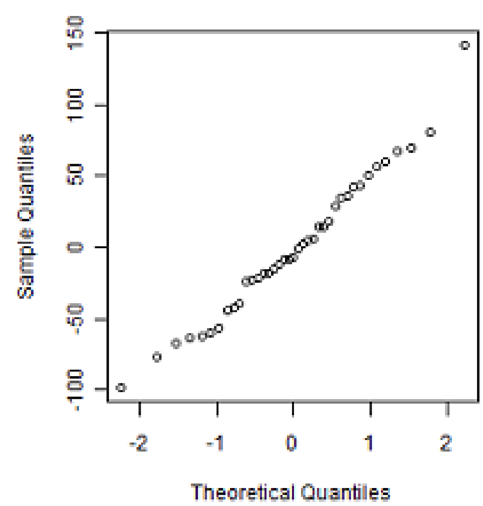
Model 2: Normal Plot

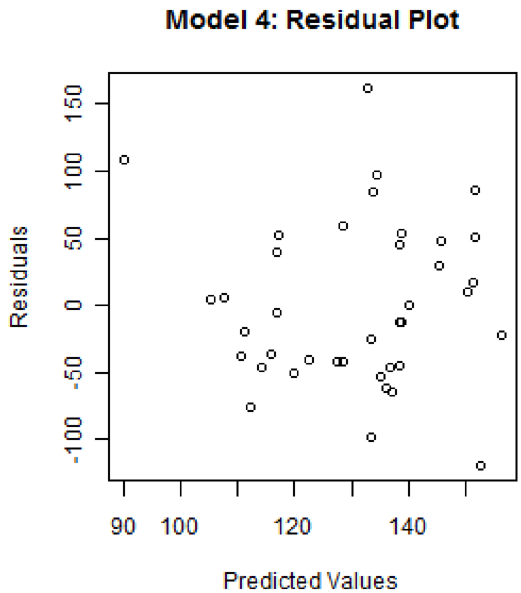


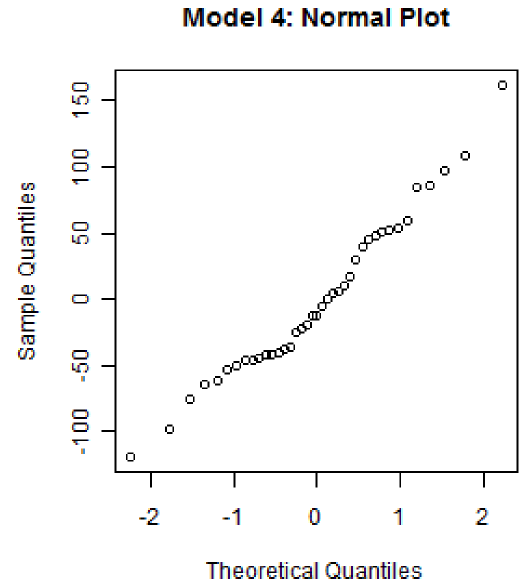
Model 3: Residual Plot

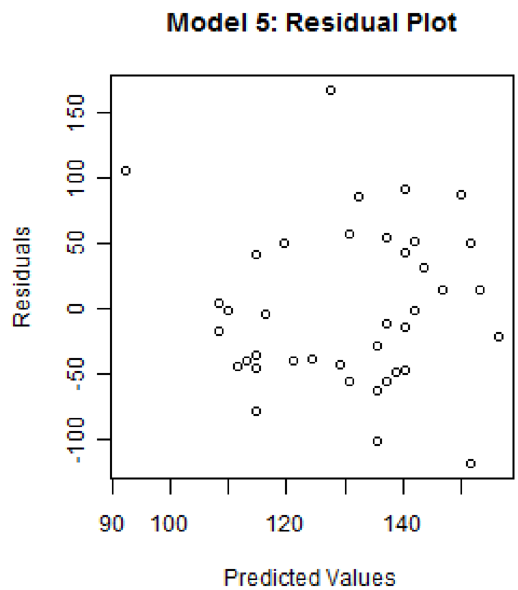


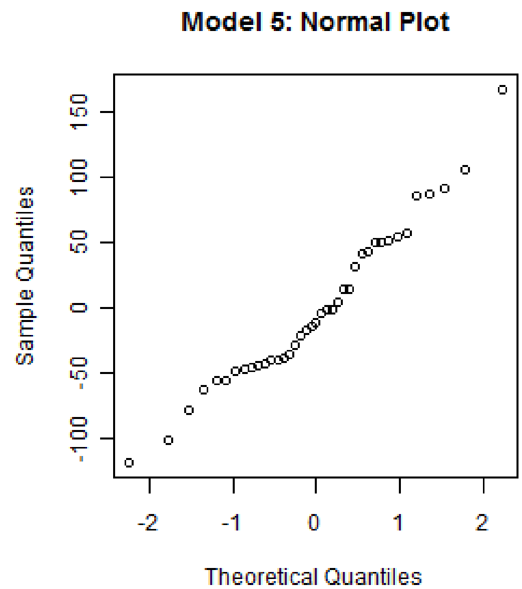
Model 3: Normal Plot

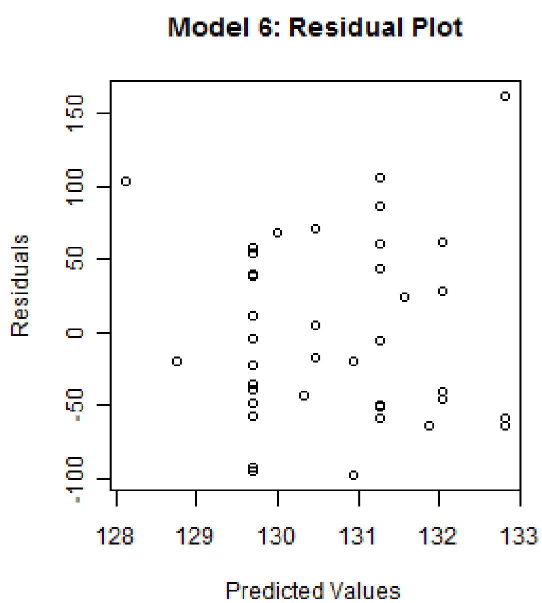


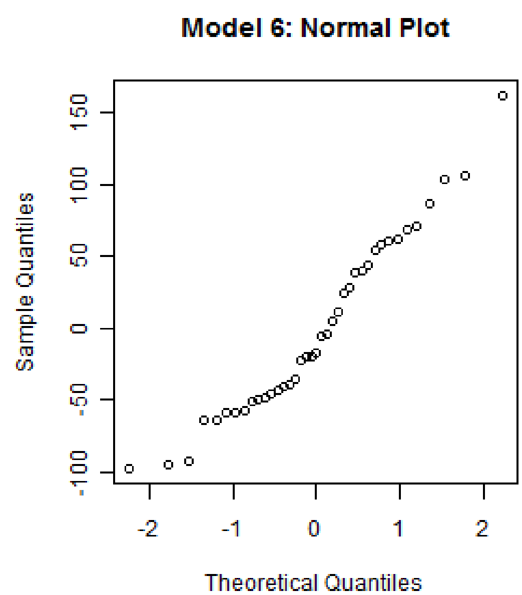

Model 4: Residual Plot
Residuals
-100
-50
0
50
100
150
90
100
120
140
Predicted Values

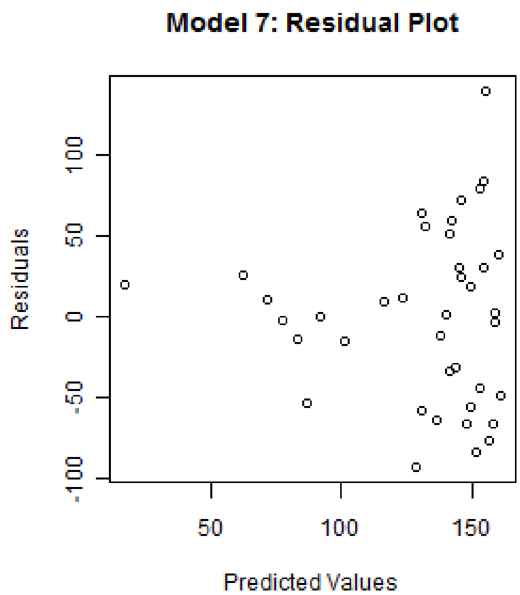

Model 4: Normal Plot
Sample Quantiles
-100
-50
0
50
100
150
-2
-1
0
1
2
Theoretical Quantiles

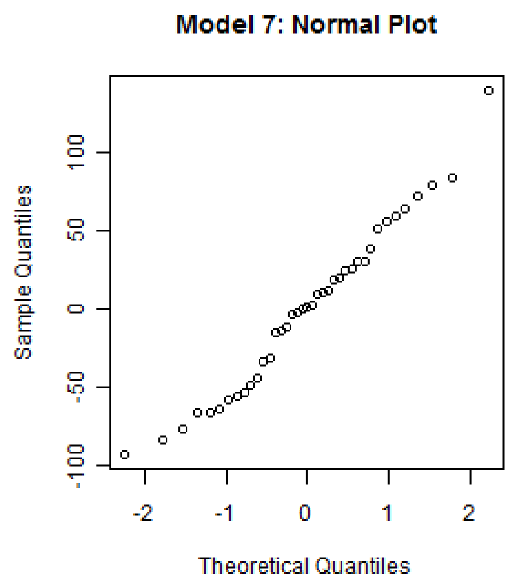

Model 5: Residual Plot
Residuals
-100
-50
0
50
100
150
90
100
120
140
Predicted Values


Model 5: Normal Plot
Sample Quantiles
-100
-50
0
50
100
150
-2
-1
0
1
2
Theoretical Quantiles


Model 6: Residual Plot
Residuals
-100
-50
0
50
100
150
128
129
130
131
132
133
Predicted Values


Model 6: Normal Plot
Sample Quantiles
-100
-50
0
50
100
150
-2
-1
0
1
2
Theoretical Quantiles


Model 7: Residual Plot
Residuals
Predicted Values
-100
-50
0
50
100
50
100
150


Model 7: Normal Plot
Sample Quantiles
Theoretical Quantiles
-100
-50
0
50
100
-2
-1
0
1
2