Test 2

STAT 021

Swarthmore College

Do not flip this page until instructed to do so.

Test organization: There are 8 questions in total on this test and they are organized into two subsections: the first **4** questions are select all that apply multiple choice questions and the last **4** questions are free response. If you need additional paper you may come to the front of the class and pick some up. There are a total of 30 points possible on this test.

Instructions: The first part of this test are multiple choice questions that do not require any additional explanation or work. No extra work will be considered in the grading of these questions but *you can get partial credit* for many of these questions. The last part of this test involves short answer questions. For these questions, you must show all your work and/or provide enough justification and explain your reasoning in order to get full credit or be considered for partial credit. You do not need a calculator to evaluate any expressions. For any calculation problems, simply writing out the formula to find the answer will suffice.

First and Last Name: _____

Swarthmore Username: _____

Take a deep breath.

You have prepared for this test and with a clear and well-rested mind, you are ready to show me what you have learned this semester. As with the other tests, the purpose of this test is to measure your understanding of the material we have covered. This is nothing more than a metric for me to evaluate your preparedness to think statistically at this particular moment in time and in this particular setting. This is not a perfect measure of your knowledge and does not predict your future statistical skills.

Multiple choice problems (2 points each)

Use the following information to answer questions 1-4.

Suppose we are trying to understand how the aerial biomass (response variable) production of a certain type of marsh grass is related to the three predictor variables

- pH, which measures the pH level of the soil (from 0-14),
- K, which measures the potassium level of the soil (in ppm), and
- the categorical variable *location* which can be one of three different spots ("OI" is short for Oak Island, "SI" is short for Smith Island, and "SM" is short for Snows Marsh).

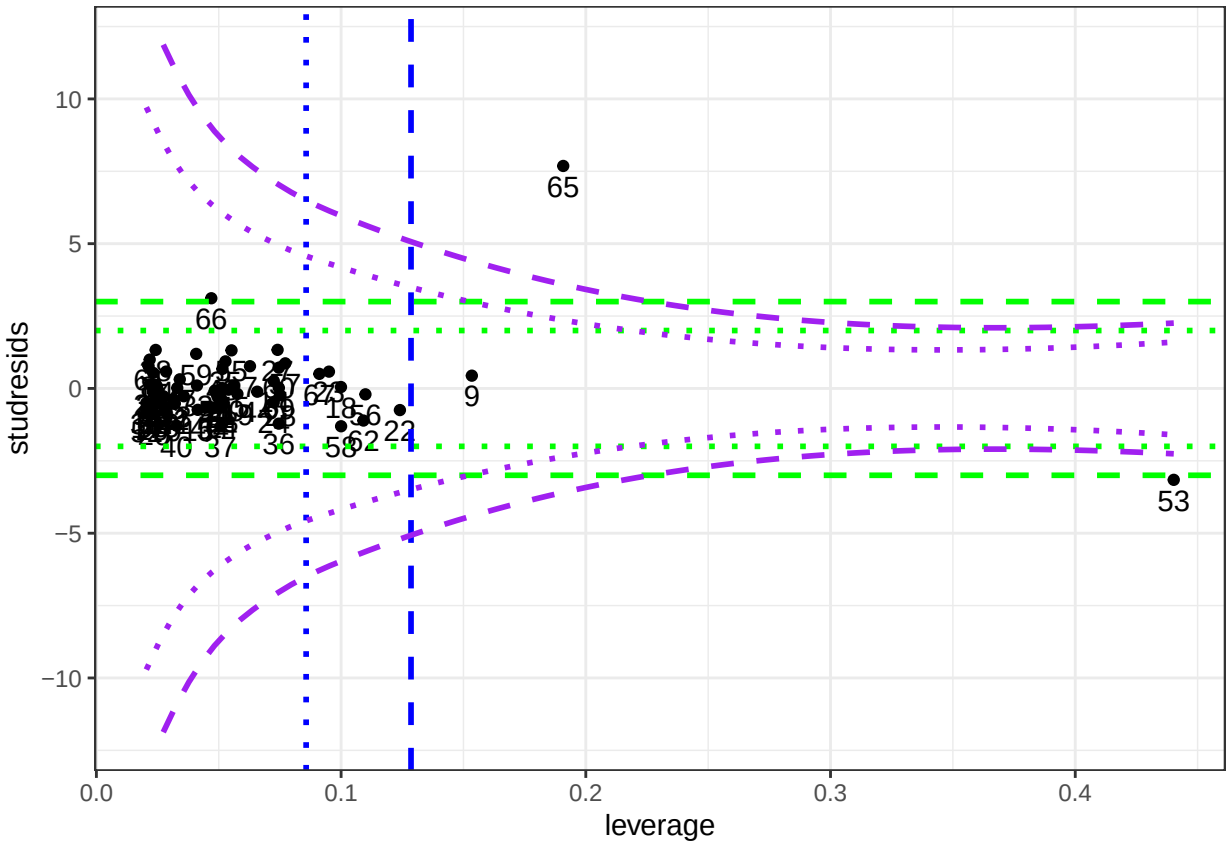
Aerial biomass is called `BIO` and is measured with units gm^{-2}.

Consider the main effects model shown below. The summary output is shown on the next page. Below the model summary is a plot of the studentized residuals and leverage values for each data point based on this model. The heavy dashed lines represent "extreme" cutoffs and the short dashed lines represent "moderate" cutoffs for leverage (blue), studentized residuals (green), and Cook's distance (purple), respectively.

$$\hat{biomass} = \hat{\beta}_0 + \hat{\beta}_1 pH + \hat{\beta}_2 K + \hat{\beta}_3 LocationSI + \hat{\beta}_4 LocationSM, \text{ where}$$

$$LocationSI = \begin{cases} 1, & \text{if at Smith Island} \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad LocationSM = \begin{cases} 1, & \text{if at Snows Marsh} \\ 0, & \text{otherwise} \end{cases}$$

```
##
## Call:
## lm(formula = BIO ~ pH + K + Location, data = biomass)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -811.15 -190.99  -37.70   96.78 1056.67
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   43.0122   299.1682   0.144   0.8864
## pH           414.9021    43.3381   9.574 6.68e-12 ***
## K             -1.0095     0.2324  -4.344 9.32e-05 ***
## LocationSI  -497.6490   163.4252  -3.045   0.0041 **
## LocationSM    58.1814   131.6870   0.442   0.6610
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 354.2 on 40 degrees of freedom
## Multiple R-squared:  0.7382, Adjusted R-squared:  0.712
## F-statistic: 28.19 on 4 and 40 DF,  p-value: 3.613e-11
```



Problem 1

Which of the following statements are supported by the R output for this model? (Circle all that apply.)

(a) Comparing grass from soil with matching pH levels and matching potassium levels, grass from the Snows Marsh tends to have higher biomass than grass from Oak Island.

(b) The effect of changing the pH level of the soil on the biomass depends on the potassium level of the soil.

(c) The effect of changing the potassium level of the soil has a greater impact on the biomass than changing the pH level, given the location of the grass is the same.

(d) The effect of changing pH level of the soil has a greater impact on the biomass than changing the potassium level, given the location of the grass is the same.

Problem 2

In comparison to the model above, which of the following are valid reduced models? (Circle all that apply.)

(a) $\widehat{biomass} = \hat{\beta}_0$

(b) $\widehat{biomass} = \hat{\beta}_0 + \hat{\beta}_1 pH + \hat{\beta}_2 K$

(c) $\widehat{biomass} = \hat{\beta}_0 + \hat{\beta}_1 pH + \hat{\beta}_2 K + \hat{\beta}_4 LocationSM$

(d) $\widehat{biomass} = \hat{\beta}_0 + \hat{\beta}_1 pH + \hat{\beta}_2 K + \hat{\beta}_3 LocationSI$

Problem 3

Which of the following data points are likely *not* very unusual with respect to their observed biomass? (Circle all that apply.)

(a) Observation 66

(b) Observation 65

(c) Observation 9

(d) Observation 53

Problem 4

Which of the following data points are likely very unusual with respect to soil pH and potassium level (K)? (Circle all that apply.)

(a) Observation 66

(b) Observation 65

(c) Observation 9

(d) Observation 53

Short answer questions

Problem 5 (6 points)

The R-output below contains the results from a regression model fit to a data set concerning life expectancy in years, alcohol consumption in drinks per day, and smoking status (smoker - Yes vs. non-smoker - No). Answer the following questions pertaining to the R output below.

```
## Call:
## lm(formula = Lifespan ~ Alcohol + Smoker)

## Residuals:
## Min 1Q Median 3Q Max
## -30.796 -7.139 0.125 6.949 19.578

## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 93.6818  2.1024     44.56   <2e-16 ***
## Alcohol     -3.2656  0.3147    -10.38   <2e-16 ***
```

```
## SmokerYes  -23.4392  1.9922    -11.77   <2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 9.959 on 97 degrees of freedom
## Multiple R-squared: 0.7129, Adjusted R-squared: 0.7069
## F-statistic: 120.4 on 2 and 97 DF, p-value: < 2.2e-16
```

a) How does drinking alcohol affect average life expectancy? (Explain in a full sentence.) (2 points)

b) Suppose someone consumes 2 alcoholic drinks per day but doesn't smoke. What is the average life expectancy for this individual? (1 point)

c) What is the average difference in lifespan between smokers and non-smokers? (1 point)

d) Suppose we consider only individuals who drink the same amount of alcohol; is there a statistically significant relationship between life expectancy and smoking habits? Justify your answer. (2 points)

For Problem 6-8 we are going to consider three different MLR models for vehicle fuel consumption (in miles per gallon) as predicted by the vehicles weight (in lbs) and possibly also by the transmission type where $transmission\_typeM = \begin{cases} 1, & \text{if manual vehicle} \\ 0, & \text{otherwise} \end{cases}$.

You can assume the vehicles were selected as a simple random sample.

Model 1: $mpg = \beta_0 + \beta_1 weight + \epsilon$

Model 2: $mpg = \beta_0 + \beta_1 weight + \beta_2 transmission\_typeM + \epsilon$

Model 3:

$mpg = \beta_0 + \beta_1 weight + \beta_2 transmission\_typeM + \beta_3 weight \cdot transmission\_typeM + \epsilon$

The summary for each model and the corresponding residual plots and Normal quantile plots are shown on last three pages after the statement of each problem.

Problem 6 (5 points)

a) Which of the three models would you choose to predict vehicle mileage? Justify your answer. (2 points)

b) State the null and alternative hypotheses for a test of a linear association among the predictors and response based on your answer to part (a). For an $\alpha = 0.05$ significance level, report the p-value and interpret the conclusion of this test in the context of the problem. (3 points)

Problem 7 (5 points)

a) State the null and alternative hypotheses for a test of the significance of the categorical predictor variable when using Model 3 as the full model. (2 points)

b) Which of the tests in Problem 6.b or Problem 7.a is most reliable? Briefly explain. (3 points)

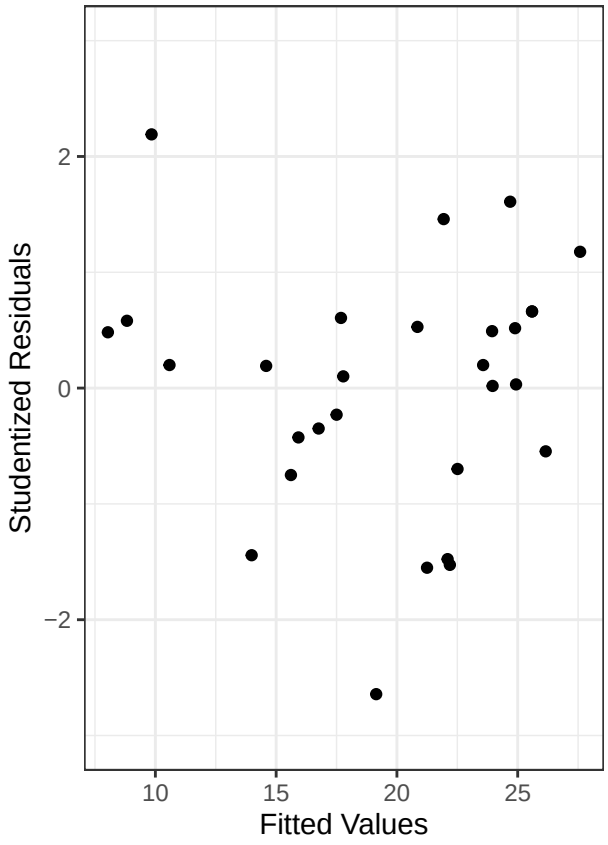
Problem 8 (6 points)

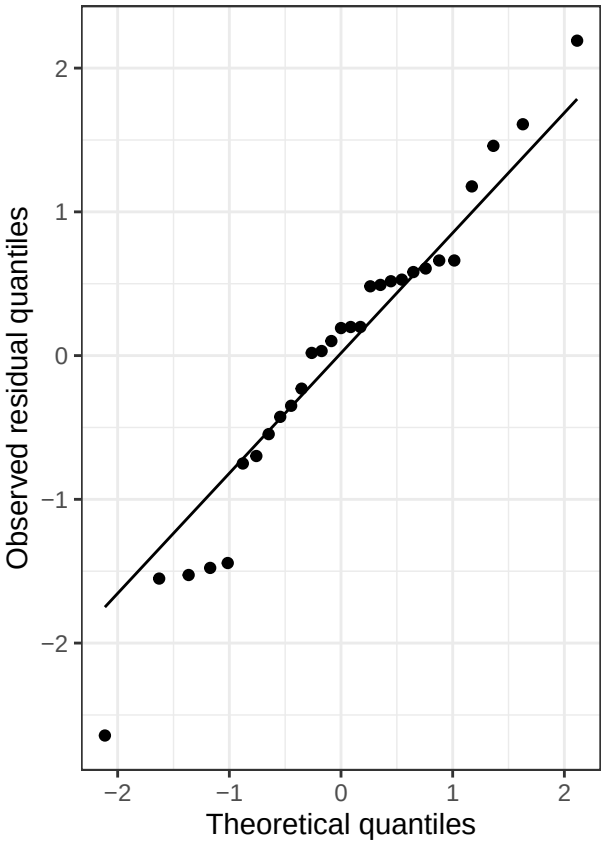
Suppose someone suggests that we should add another predictor variable displacement, which measures the displacement of the vehicle's engine (in inches). Describe what steps you could take to statistically support (or not) this decision without conducting any tests or calculating any confidence intervals. (More points will be awarded for more valid methods with accurate justifications.)

Model 1:

$mpg = \beta_0 + \beta_1 weight + \epsilon$

```
##
## Call:
## lm(formula = mpg ~ weight, data = car_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1418 -1.1597  0.4131  1.1759  4.1569
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.9659944  1.5594404   25.63  < 2e-16 ***
## weight      -0.0067480  0.0004985  -13.54  1.5e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.189 on 27 degrees of freedom
## Multiple R-squared:  0.8716, Adjusted R-squared:  0.8668
## F-statistic: 183.2 on 1 and 27 DF,  p-value: 1.502e-13
```



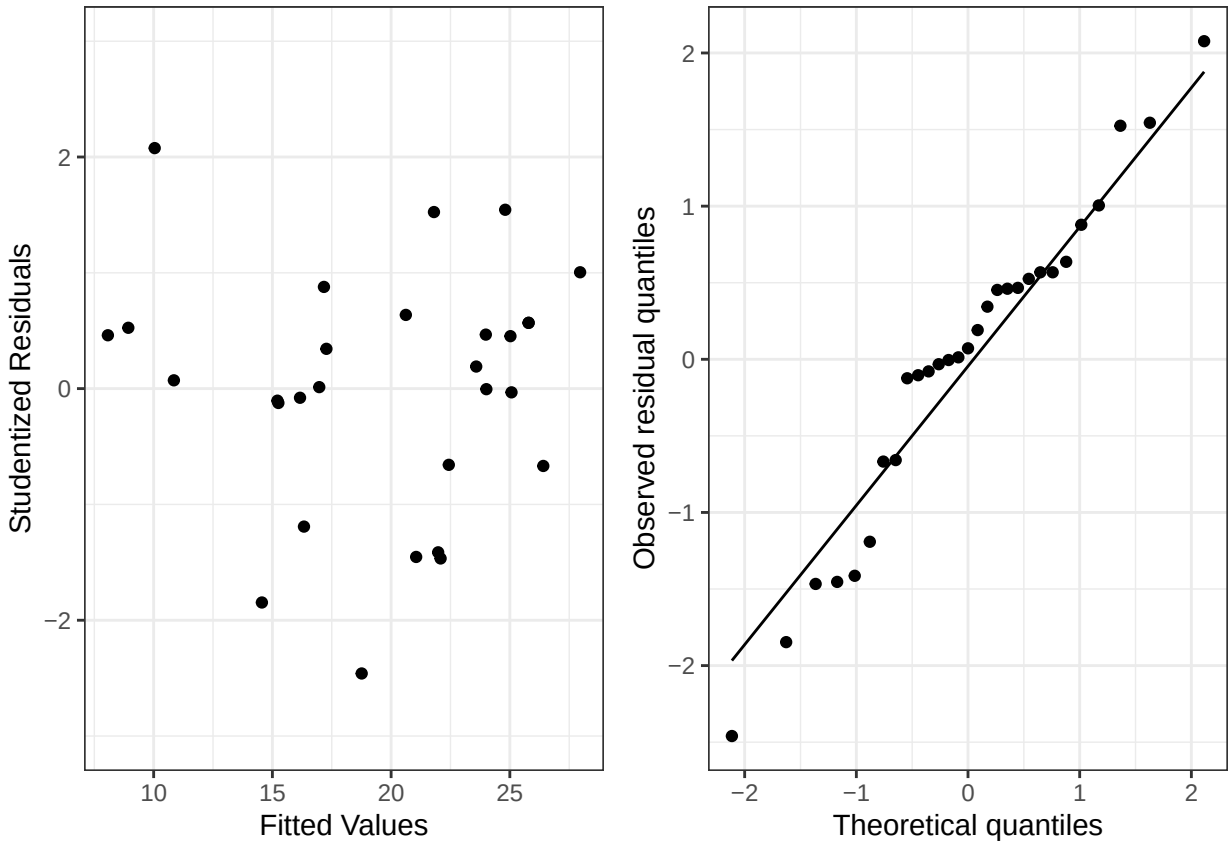


.

Model 2:

$mpg = \beta_0 + \beta_1 weight + \beta_2 transmission\_type M + \epsilon$

```
##
## Call:
## lm(formula = mpg ~ weight + transmission_type, data = car_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7600 -1.4079  0.1482  1.2099  3.9571
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         41.4534530  2.3336110   17.76 4.63e-16 ***
## weight              -0.0073537  0.0008641   -8.51 5.45e-09 ***
## transmission_typeM   1.4161833  1.6463688    0.86    0.398
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.199 on 26 degrees of freedom
## Multiple R-squared:  0.8751, Adjusted R-squared:  0.8655
## F-statistic: 91.11 on 2 and 26 DF,  p-value: 1.795e-12
```



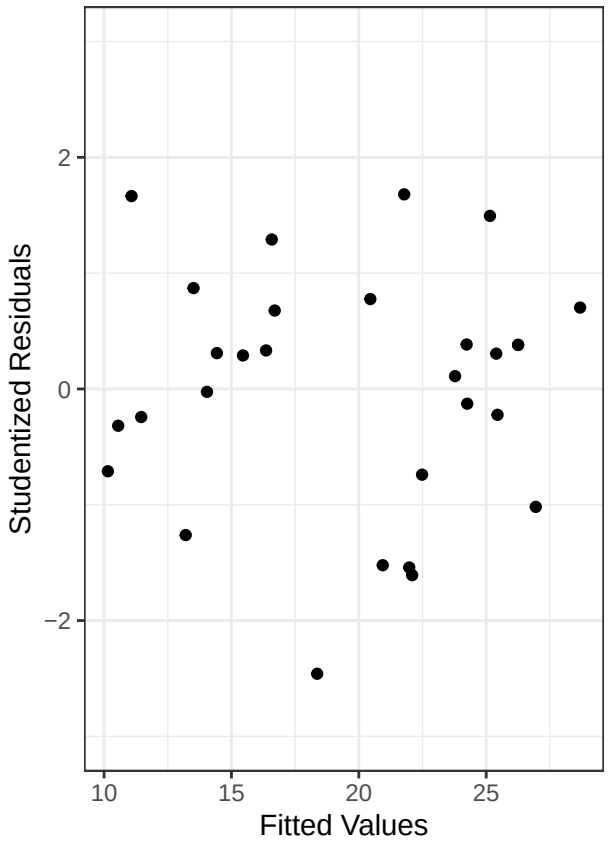
.

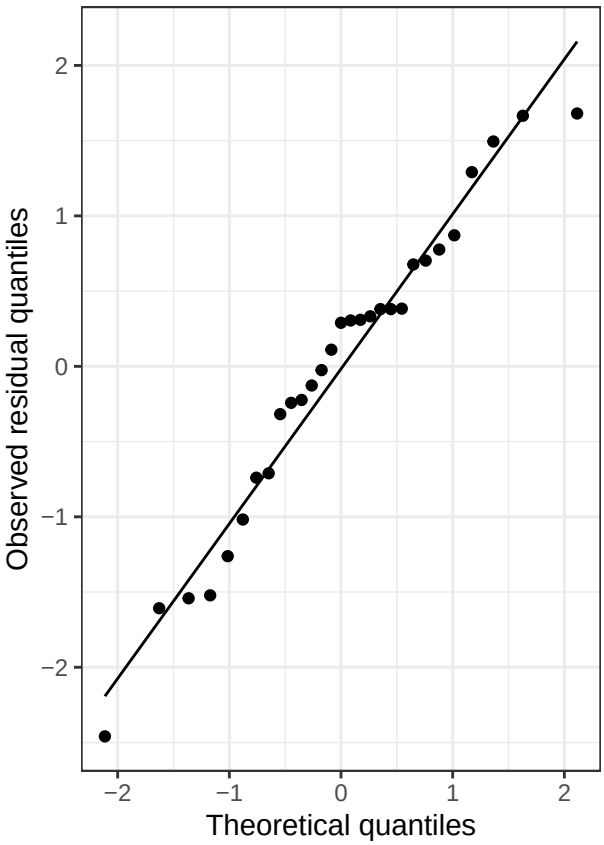
Model 3:

$mpg = \beta_0 + \beta_1 weight + \beta_2 transmission\_typeM + \beta_3 weight \cdot transmission\_typeM + \epsilon$

```
##
## Call:
## lm(formula = mpg ~ weight + transmission_type + weight:transmission_type,
##     data = car_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3650 -1.1503  0.5474  1.3016  3.2194
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                4.383e+01  2.382e+00  18.400 4.79e-16 ***
## weight                    -8.250e-03  8.854e-04  -9.318 1.31e-09 ***
## transmission_typeM        -1.729e+01  8.136e+00  -2.125   0.0437 *
## weight:transmission_typeM  4.787e-03  2.046e-03   2.340   0.0276 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.031 on 25 degrees of freedom
## Multiple R-squared:  0.8976, Adjusted R-squared:  0.8853
## F-statistic: 73.02 on 3 and 25 DF,  p-value: 1.669e-12
```

Model 3:


Studentized Residuals
Fitted Values
2
0
−2
10
15
20
25


Observed residual quantiles
Theoretical quantiles
2
1
0
−1
−2
−2
−1
0
1
2