ISLR Chap 3 Notes

M. E. Waggoner

February 3, 2019

Contents

3 Simple Linear Regression

Note: The section numbers of this document match up with ISLR. The equation numbers match up with the section, but you need to append "3." to the number.

The statistics behind most machine learning assumes that the data we have is just one sample of a much larger population. Thus, when we calculate means and standard deviations, we are calculating them from a sample, i.e.,

$$\bar{x} = \frac{1}{n}\Sigma_{i-1}^{n} x_i.$$

Sample statistics methods are what are behind the hypothesis testing and confidence intervals we will use.

y is aproximately a linear function of X, but unknown

$$Y \approx \beta_0 + \beta_1 X \tag{1}$$

Since we don't know the slope and intercept, we find approximations for them using the training data.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2}$$

3.1 Estimating the Coefficients

ith Residual. Note that this is the difference between the actual value and the estimated linear relationship, which is not the same as ϵ_i, which is the difference between the actual value and the true line.

$$e_i = y_i - \hat{y}_i$$

Residual sum of squares = what is minimized by least squares. I said that MSE was minimized by least squares, and it is since $MSE = RSS/n$.

$$\begin{aligned}\text{RSS} =& e_1^2 + e_2^2 + \cdots + e_n^2 \\ =& (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2 \\ =& \Sigma_{i-1}^{n} (y_i - \hat{y}_i)^2\end{aligned} \tag{3}$$

Least squares coefficients estimates that minimize RSS where $\bar{x}$ and $\bar{y}$ as calculated by `lm()`.

$$\begin{aligned}\hat{\beta}_1 =& \frac{\Sigma_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\Sigma_{i=1}^{n}\left(x_i-\bar{x}\right)^2} \\ \hat{\beta}_0 =& \bar{y}-\hat{\beta}_1\bar{x},\end{aligned} \tag{4}$$

where $\bar{x}$ and $\bar{y}$ are means of x and y, respectively.

3.2 Assessing the Accuracy of the Coefficient Estimates

These statistics assess the accuracy of the parameters of slope and intercept, and do not assess the accuracy of the model. In other words, these statistics are evaluating our ability to use the estimates of the slope and intercept, but they do not tell us whether a linear model is a good fit to the data.

True linear relationship between X and Y

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{5}$$

The standard error (SE) of a statistic (usually an estimate of a parameter) is the standard deviation of its sampling distribution or an estimate of that standard deviation.

Standard error of a sample mean where $\sigma^2 = \mathrm{Var}(x)$ for a population X

$$\mathrm{Var}\left(\hat{\mu}\right) = \mathrm{SE}\left(\hat{\mu}\right)^2 = \frac{\sigma^2}{n} \tag{7}$$

Standard error of estimates of linear coefficients where $\sigma^2 = \mathrm{Var}(\epsilon)$ are calculated by `lm()`.

$$\begin{aligned}\mathrm{SE}\left(\hat{\beta}_0\right)^2 &= \sigma^2\left[\frac{1}{n}+\frac{\bar{x}^2}{\Sigma_{i-1}^{n}\left(x_i-\bar{x}^2\right)}\right] \\ \mathrm{SE}\left(\hat{\beta}_1\right)^2 &= \frac{\sigma x^2}{\Sigma_{i-1}^{n}\left(x_i-\bar{x}^2\right)}\end{aligned} \tag{8}$$

We will use the standard errors in two ways. In practice, usually one or the other is used.

3.2.1 Confidence intervals

There is approximately a 95% chance that the true value of the slope β_1 lies in the interval

$$\left[\hat{\beta}_1 - 2\mathrm{SE}\left(\hat{\beta}_1\right), \hat{\beta}_1 + 2\mathrm{SE}\left(\hat{\beta}_1\right)\right]. \tag{10}$$

If the confidence interval for the slope contains 0, this is an indication that the slope very well might be 0 and that Y does not depend linearly on X.

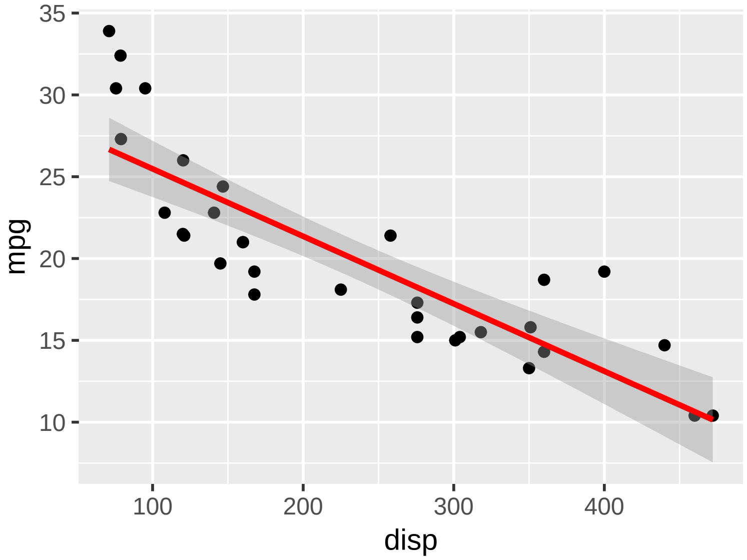
There is approximately a 95% chance that the true value of the intercept β_0 lies in the interval

$$\left[\hat{\beta}_0 - 2\text{SE}\left(\hat{\beta}_0\right), \hat{\beta}_0 + 2\text{SE}\left(\hat{\beta}_0\right)\right]. \tag{11}$$

Generally, we will use a confidence level of 95%, but in practice the confidence level will be determined by the application, the client, and the industry.

3.2.1.1 Example: How large is the effect?

Consider the `mtcars` dataset that comes with base R. Is the miles per gallon `mpg` a car can get a linear relationship of `disp`? If so, how large is the effect? That is, what is the change in `disp` for an increase or decrease of 1 `mpg`?



We can view the coefficients and their standard errors using the `tidy()` function.

term	estimate	std.error	statistic	p.value
(Intercept)	29.5999	1.2297	24.070	0
disp	-0.0412	0.0047	-8.747	0

The estimate of the slope is $\hat{\beta}_1 = -0.0412$ with $\text{SE}\left(\hat{\beta}_1\right) = 0.0047$. Thus, there is a 95% chance that the true slope is in the interval $[-0.0508, -0.0316]$. This interval is relatively narrow since $\frac{\hat{\beta}_1}{\hat{\beta}_1} = -0.1143$ is small. This interval is "far" from zero since it does not contain zero and the distance from the interval to zero relatively large since $\left|\frac{\hat{\beta}_1 + \text{SE}(\hat{\beta}_1)}{\text{SE}(\hat{\beta}_1)}\right| = 7.7472$ is large.

The estimate of the intercept is $\hat{\beta}_0 = 29.5999$ with $\text{SE}\left(\hat{\beta}_0\right) = 1.2297$. Thus, there is a 95% chance that the true intercept is in the interval $[27.0884, 32.1113]$.

This information can be found with the `confint()` function.

	2.5 %	97.5 %
(Intercept)	27.0884	32.1113
disp	-0.0508	-0.0316

So how large is the effect of `mpg` on `disp`? The change in `disp` given a change in `mpg` is

$$\frac{dy}{dx} = \frac{d}{dx}\left(\beta_1 x + \beta_0\right) = \beta_1.$$

Thus, there is a change of -0.0417 ± 0.0047 miles per gallon for each change of 1 unit of displacement.

3.2.2 Hypothesis testing

A common hypothesis test of linear relationships tests the null hypothesis

$$\begin{aligned} &H_0 : \text{There is no relationship between } X \text{ and } Y \\ &H_0 : \beta_1 = 0 \end{aligned} \tag{12}$$

versus the alternate hypothesis

$$\begin{aligned} &H_1 : \text{There is a relationship between } X \text{ and } Y \\ &H_1 : \beta_1 \neq 0 \end{aligned} \tag{13}$$

using the t-statistic

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}\left(\hat{\beta}_1\right)} \tag{14}$$

which measures the number of standard deviations that $\hat{\beta}_1$ is away from the mean 0, because if there is no relationship between X and Y, then t will have a t-distribution with $df = n - 2$, that is, the degrees of freedom = the number of observations − the number of parameters.

3.2.2.1 Example: Is there a relationship?

Continuing the example of `mtcars`, we can ask if there a relationship between miles per gallon and engine displacement. For simple linear regression, this question can be answered with some level of confidence using a hypothesis test. The null hypothesis is that there is no relationship, and it is assumed until we have evidence to the contrary. The alternate hypothesis is what we will believe if there is statistical evidence.

$$\begin{aligned} &H_0 : \text{There is no relationship between mpg and displacement or } \beta_1 = 0 \\ &H_1 : \text{There is a relationship between mpg and displacement or } \beta_1 \neq 0 \end{aligned}$$

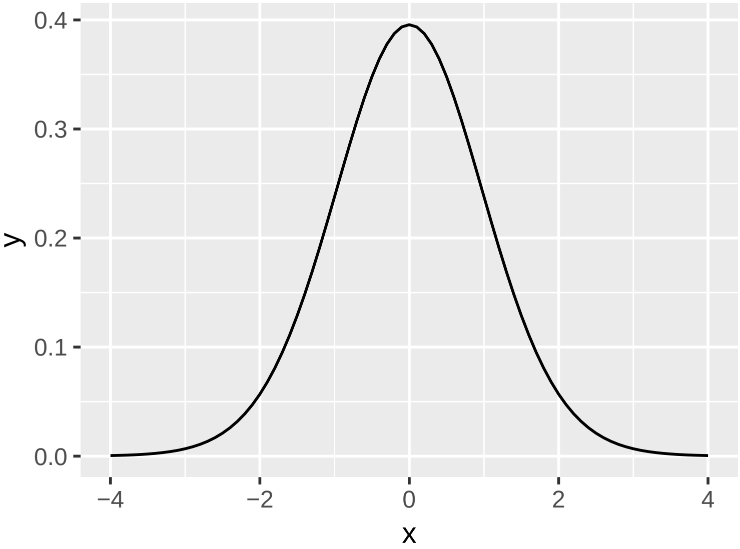
We'll illustrate in another lab that the distribution of all possible slopes is approximately the t-distribution with $n - 2$ degrees of freedom. The t statistic for the slope for this model is

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}\left(\hat{\beta}_1\right)} = \frac{-0.0412}{0.0047} = -8.7472.$$

At a significance level of 5%, we will reject the null hypothesis if

$$p = P(|t| > |-8.747|) < 5\%.$$

To put this in perspective, here is the plot of the t-distribution for df $= 30$. I think -8.747 is somewhere in the next room.



From the output of `lm()`, we see that $p = 0 < 5\%$. Therefore, at the 5% significance level we can conclude that there is a relationship between miles per gallon and engine displacement. At this point we don't know if we have modeled that relationship well or not.

We can ask the same thing about the `mpg`-intercept: is there evidence that the intercept is not zero? The null hypothesis is that the intercept is zero, and it is assumed until we have evidence to the contrary. The alternate hypothesis is what we will believe if there is statistical evidence.

$$\begin{aligned} H_0 &: \text{The intercept is zero or } \beta_0 = 0 \\ H_1 &: \text{The intercept is not zero or } \beta_0 \neq 0 \end{aligned}$$

We can demonstrate that the distribution of all possible intercepts is approximately the t-distribution with $n - 2 = 30$ degrees of freedom for `mtcars`. The t statistic for the intercept for this model is

$$t = \frac{\hat{\beta}_0 - 0}{\text{SE}\left(\hat{\beta}_0\right)} = \frac{29.5999}{1.2297} = 24.0704.$$

At a significance level of 5%, we will reject the null hypothesis if the area beyond this t-statistic is less than 5%. That is, we reject the null hypothesis if

$$p = P(|t| > |24.07|) < 5\%.$$

The t-distribution for the intercept is the same as for the slope. From the output of `lm()`, we see that $p = 0 < 5\%$ for the intercept. Therefore, at the 5% significance level we can conclude that the `mpg`-intercept is not zero. However, we do not know yet whether we have made a good estimate of the intercept or not.

3.3 Assessing the Accuracy of the Model

These statistics allow us to assess the accuracy of the model; that is, they help us know whether a linear model was a good choice.

3.3.1 Residual Standard Error

A small RSE *relative to the data* indicates that the actual data are close to the predictions, and we can say the model fits the data well.

A large RSE *relative to the data* indicates that the actual data are not close to the predictions, and we can say the model does not fit the data well.

The *residual standard error* is

$$
\begin{aligned}
\mathrm{RSE} &= \sqrt{\frac{1}{n-2}\mathrm{RSS}} \\
&= \sqrt{\frac{1}{n-2}\Sigma_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2} \\
&\approx \mathrm{Var}(\epsilon)^2
\end{aligned}
\tag{15}
$$

where the *residual sum of squares* is

$$
\mathrm{RSS} = \Sigma_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2. \tag{16}
$$

3.3.2 R^2 Statistic

The formula for R^2 is

$$
R^2 = \frac{\mathrm{TSS}-\mathrm{RSS}}{\mathrm{RSS}}, \tag{17}
$$

where the *total sum of squares* is

$$
\mathrm{TSS} = \Sigma_{i=1}^{n}\left(y_i-\bar{y}\right)^2.
$$

The *correlation* of X and Y is

$$
r = \mathrm{Cor}(X,Y) = \frac{\Sigma_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sqrt{\Sigma_{i=1}^{n}\left(x_i-\bar{x}\right)}\sqrt{\Sigma_{i=1}^{n}\left(y_i-\bar{y}\right)}} \tag{18}
$$

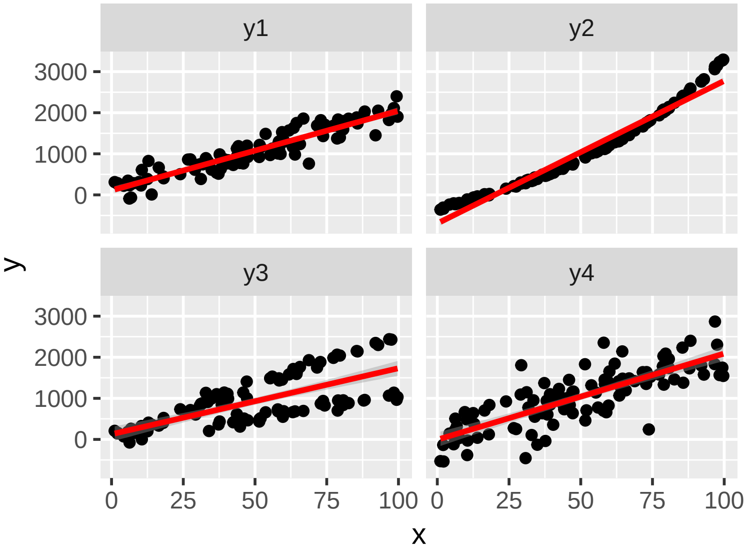
and in simple linear regression

$$
r^2 = R^2.
$$

3.3.2.1 Examples: RSE versus R^2

Or, Residual Squared Error vs Coefficient of Determination.

Consider this simulated data.



```
## # A tibble: 1 x 4
##   meany1 meany2 meany3 meany4
##    <dbl>  <dbl>  <dbl>  <dbl>
## 1  1068.  1022.   921.  1030.
```

```
##
## ============================================================
##                Model 1      Model 2       Model 3      Model 4
## ------------------------------------------------------------
## (Intercept)   110.02 *     -694.96 ***   126.21        -7.18
##               (42.55)       (41.60)      (89.80)      (89.89)
## x              19.37 ***     34.72 ***    16.06 ***    20.98 ***
##                (0.75)        (0.74)       (1.59)       (1.59)
## ------------------------------------------------------------
## R^2             0.87          0.96         0.51         0.64
## Adj. R^2        0.87          0.96         0.50         0.64
## Num. obs.     100           100          100          100
## RMSE          204.78        200.23       432.20       432.63
## ============================================================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

3.3.2.2 Example: How strong (i.e., accurate) is the relationship?

Let's return to the `mtcars` example, for which some statistics are listed below.

Using residual error, we see that $\mu\,(\text{disp}) = 230.7219$ while the RSE $= 3.2515$. Thus, there is a relative error of roughly $\frac{\text{RSE}}{\mu(\text{disp})} = 1.4093\%$. This seems small, but whether it is small enough depends on the application and the context.

The *coefficient of determination* is $R^2 = 0.72$. This says that 72% of the variation in miles per gallon is explained by this linear model and that 28% of the variation is unexplained by the model, and must either be considered error or the model needs to be changed to account for this variation.

```
##
## ======================
##                Model 1
## ----------------------
## (Intercept)  29.60 ***
##              (1.23)
## disp         -0.04 ***
##              (0.00)
## ----------------------
## R^2           0.72
## Adj. R^2      0.71
## Num. obs.    32
## RMSE          3.25
## ======================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

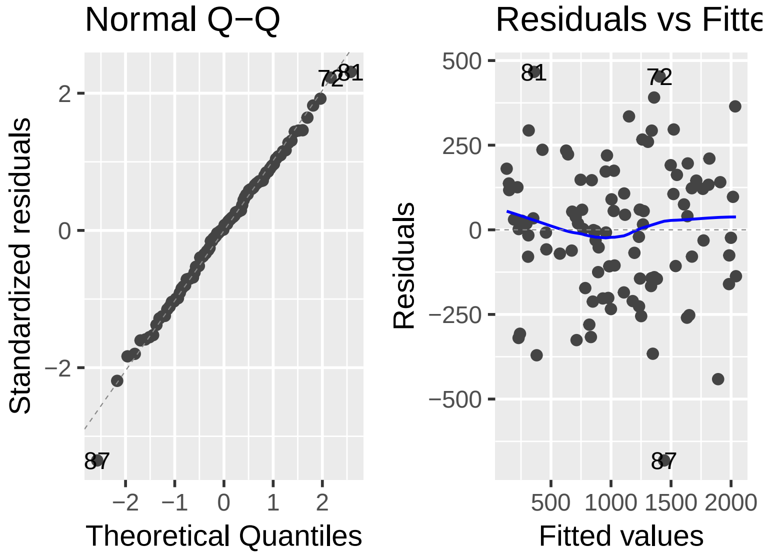
3.4 Looking to the future: Section 3.3.3

As we can see from the previous section, the values of R^2 and RSE do not tell the whole story. We would look at a variety of analyses to help us understand the structure of the data and find a good model.

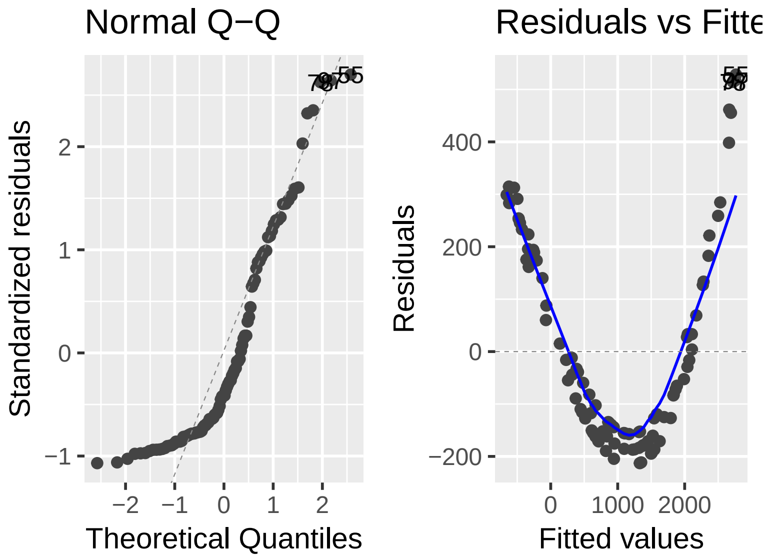
One such analysis is to look at the residuals. If our model accounts for all variation in Y that is based on X, then the residuals e_i should be a good estimate for the errors ϵ_i. Thus, the residuals should be random with a normal distribution and have not other structure. If they do not have these qualities, then there is more to the relationship of X and Y than we accounted for in the regression.

There are many ways to look the distribution of the residuals. Here we look at their distribution as a scatterplot and a normal QQ plot. These are created automatically when we call `lm()`. The command `ggfortify::autoplot()` is one way to view the data, but there are many others.

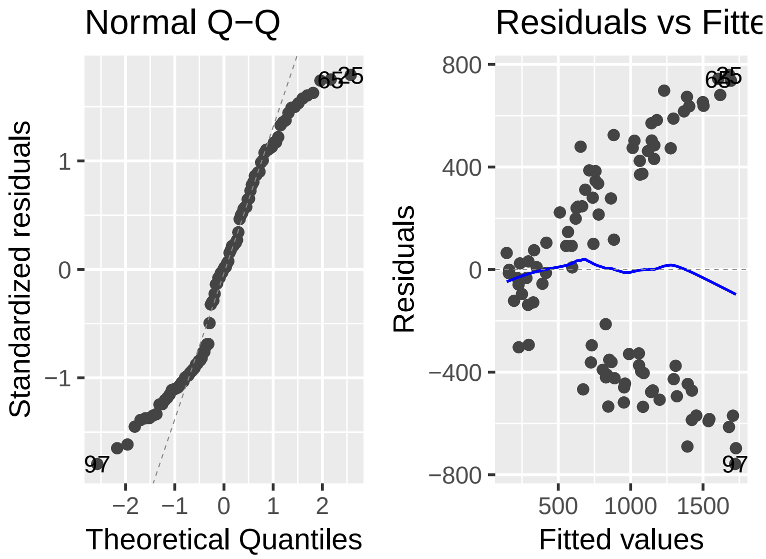
The first model `lm1` shows exactly what the residuals should look like if the linear model captured all the variation in Y that depends on X. The QQ plot shows that the distribution is normal. The scatterplot of the residuals shows that about 2/3 of the points are within one standard deviation of the mean, and almost all are within 2 standard deviations. The blue fitted line should be the line $e = 0$ if the residuals are centered around 0, and we see here that it is close.



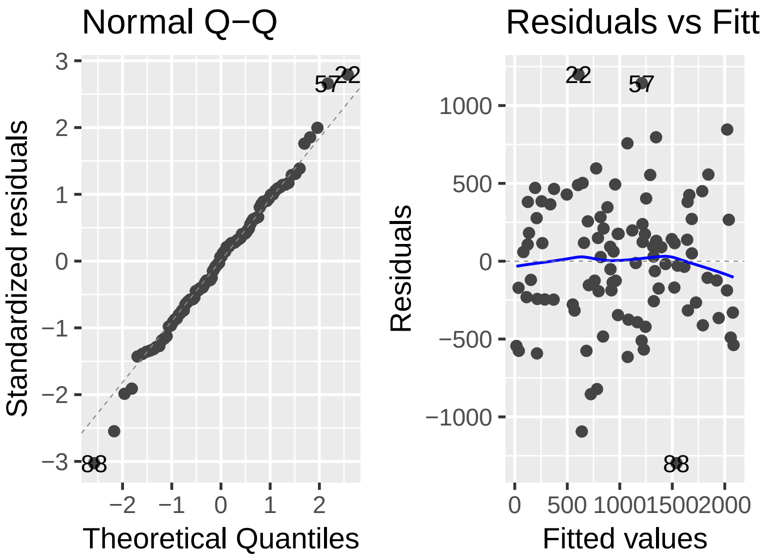
The second model `lm2` shows poor results in the residual analysis, and there is clearly a relationship between X and Y that was not capture by the linear model. The QQ plot shows that the residuals are not normally distributed. The scatterplot of the residuals has a clear U-shaped pattern, and the residuals are not random. In this case, it would be good to try a quadratic rather than linear model.



The third model `lm3` also shows there is a relationship between X and Y that was not capture by the linear model. This time the scatterplot of the residuals has a sideways funnel shape, but the blue fitted line still hugs the horizontal axis. In this case, it would be good to a Box-Cox transformation, possibly a log transformation.



The fourth model `lm4` is similar to the first model: the residuals are random and normally distributed.


Normal Q−Q
Standardized residuals
Theoretical Quantiles
3
2
1
0
−1
−2
−3
−2
−1
0
1
2
22
57
88
Residuals vs Fitt
Residuals
Fitted values
1000
500
0
−500
−1000
0
500
1000
1500
2000
22
57
88