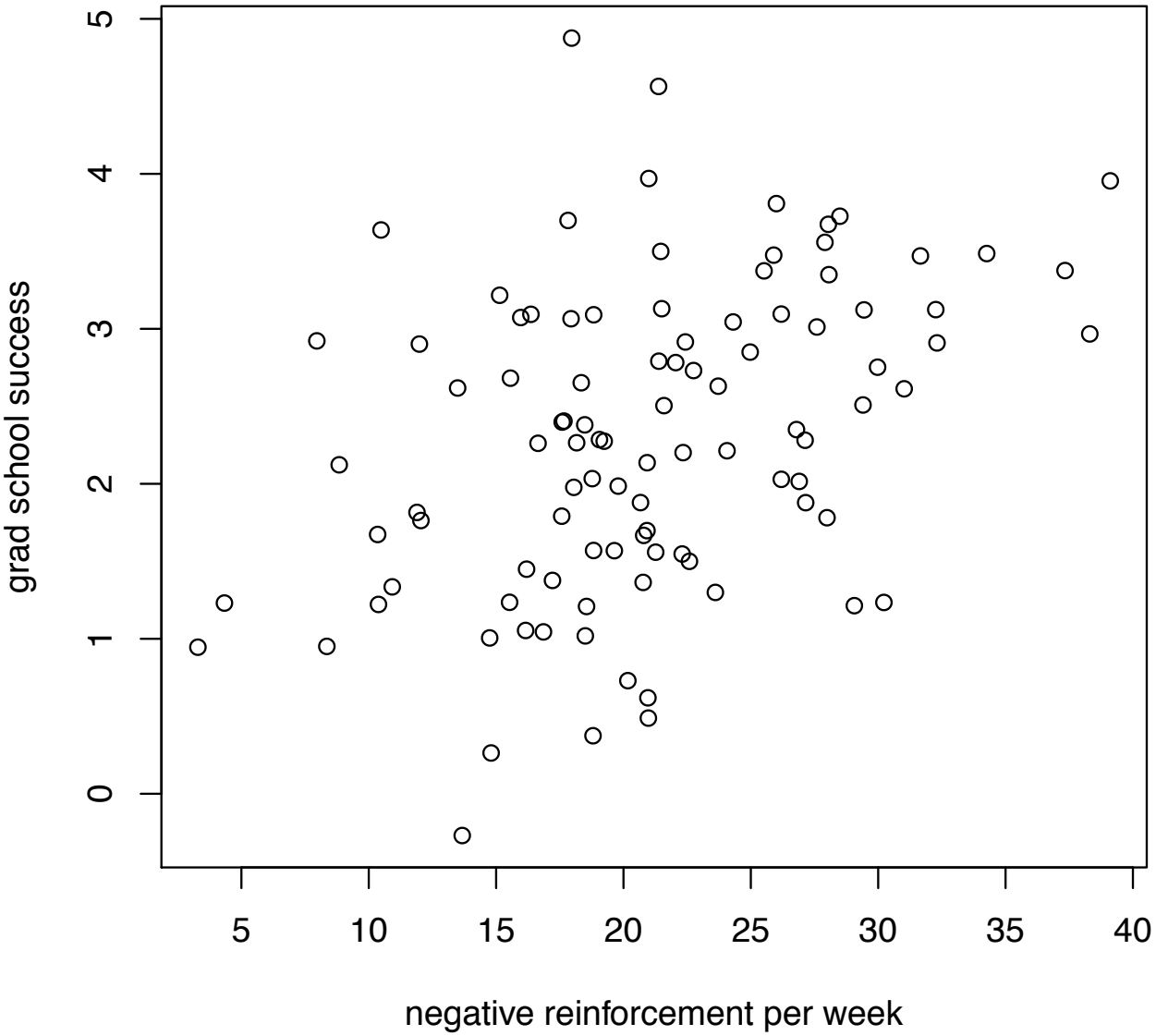
Statistics and Linguistic Applications
Hale
March 3, 2010

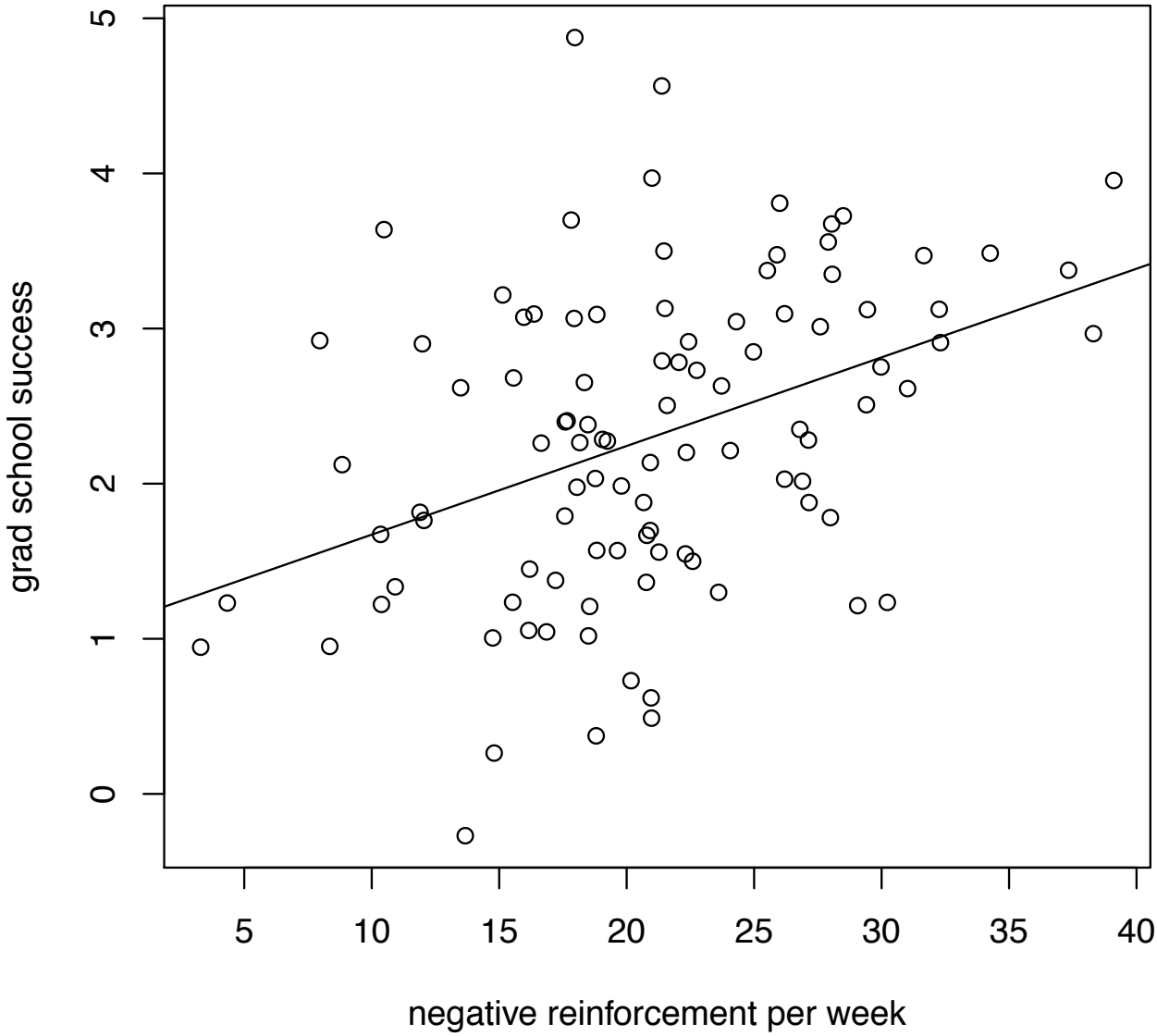
Draw a line

Pretend that these points (x, y) represent measurements where x is the amount of negative reinforcement per week from the advisor (maybe epithets per week) and y is graduate school success (perhaps quantified in number of job offers). Do your best to draw a straight line that will enable you to predict grad-school success from amount of negative (verbal) reinforcement.



DO NOT TURN THE PAGE UNTIL YOU HAVE MARKED YOUR LINE!
NO CHEATING
THIS MEANS YOU

Did you get this line?

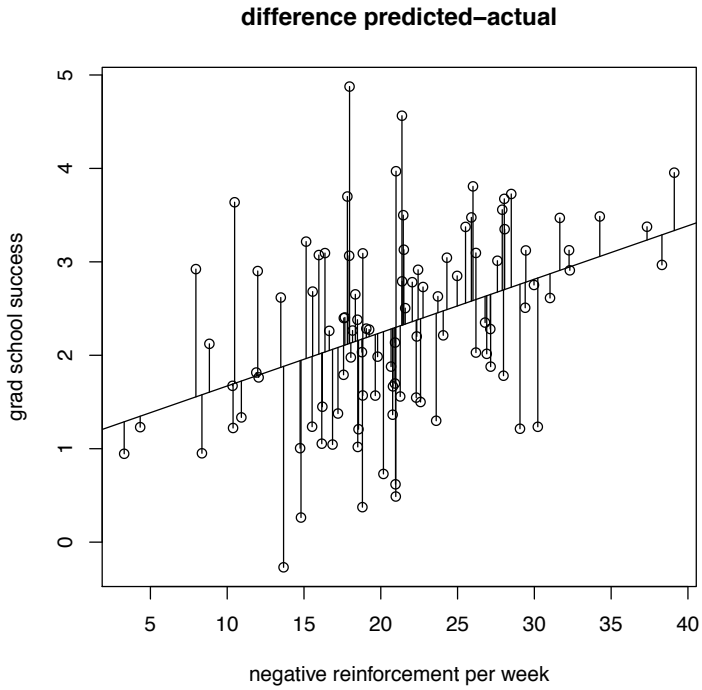


This line, like all straight lines, can be described by an equation. In this case the equation is 1.

$$\begin{aligned} y &= \text{slope} \times x + \text{intercept} \\ y &= 0.057x + 1.1 \end{aligned} \tag{1}$$

For every weekly epithet, the number of job offers after grad school increases by just over one-twentieth. Even if your advisor hurls no epithets, you still get 1.1 job offers.

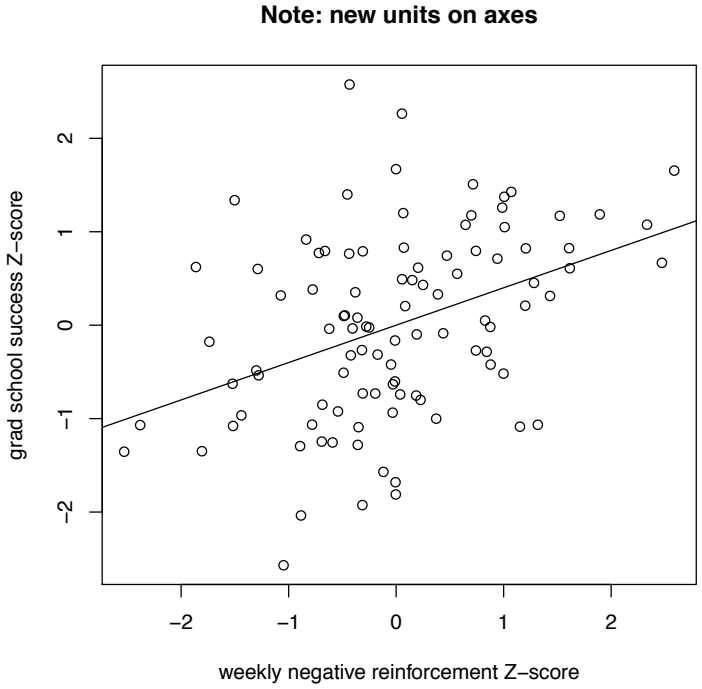
On what measure is this is "best" line? It is the line the minimizes the squared error. Each error is the squared distance between $y_{predicted}$, the degree of grad school success that the equation predicts for a given level of verbal abuse, and y the actual degree of success.



The **regression** line is the line which minimizes error. This "least squares" definition of error is due to C.F. Gauss and quite traditional (although it is possible to define it otherwise).

Lets standardize all of our measurements by converting them to Z-scores:

```
> stdx <- scale(x)
> stdy <- scale(y)
```



The standardized line has the equation

$$y = 0.4x + 0.0 \tag{2}$$

Interestingly, the correlation coefficient r in the original data is *also* 0.4, just like the slope in equation 2.

```
> cor(x, y)

[1] 0.4
```

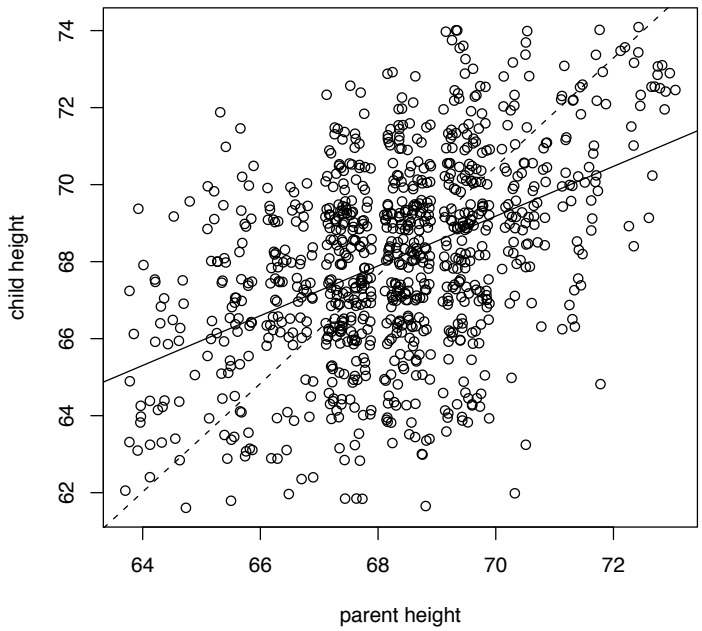
This is not a coincidence. In general, r will convert a standardized x-score into a (predicted) standardized y-score. To talk about regular old unstandarized scores it's necessary to remove ("undo") the standarization (Vasishth section 7.4.3). The slope of the regression line is thus $r \times \frac{s_y}{s_x}$.

```
> syBysx <- sd(y)/sd(x)
> cor(x, y) * syBysx

[1] 0.05714286

> coefficients(gradsuccess)

(Intercept)           x
 1.10000000  0.05714286
```

Why it is called regression

It was Francis Galton who in 1885 looked at measurements of the heights of fathers and sons.

```
> j <- 4
> plot(jitter(parent, factor = j), jitter(child, factor = j), xlab = "parent height",
+     ylab = "child height")
> riserun <- sd(child)/sd(parent)
> intercept <- mean(child) - (riserun * mean(parent))
> abline(intercept, riserun, lty = "dashed")
> regress <- lm(child ~ parent)
> abline(regress)
```



You might think that taller parents would consistently have taller children. That is, you might think that going up by one standard deviation in parent-height would always yield a corresponding one standard deviation in child height. The dashed line expressed this (false) idea . Although taller parents do tend to have

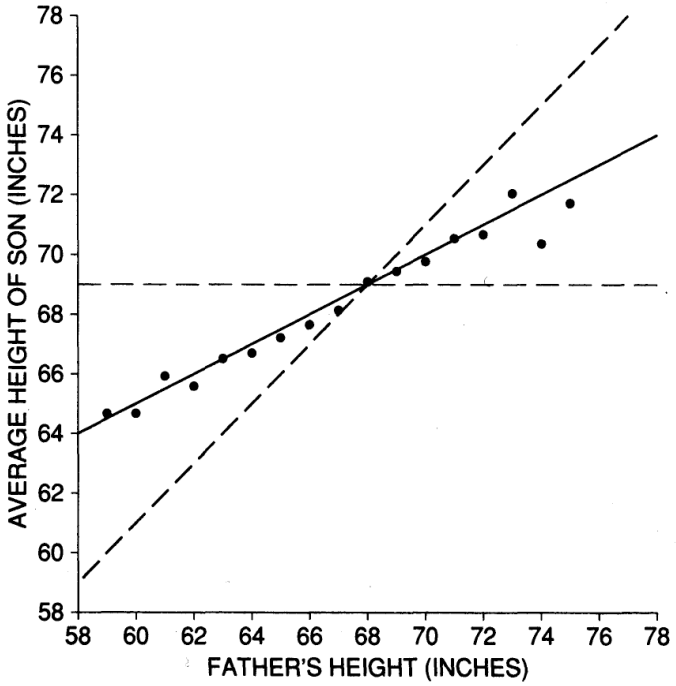


Figure 1: Galton saw the regression fallacy

taller children, extremely tall parents typically have children whose heights go back down or "regress toward mediocrity." That is, the children's heights are more like the average. At the same time, extremely short parents' kids tend to be a little taller, again regressing toward mediocrity.

```
> cbind(mean(parent), mean(child), mean(child[parent == 64]), mean(child[parent ==
+     73]))

         [,1]     [,2]     [,3]  [,4]
[1,] 68.30819 68.08847 65.30714 72.95
```

The cloud in general is symmetric about the dashed $\sigma_{child}/\sigma_{parent}$ line, but only on average — not at particular vertical strips. The solid regression line optimizes for each vertical strip in an attempt to minimize all deviations, not just deviations from the centroid. The difference between the 1SD/1SD line and the regression line is even starker in figure 1.

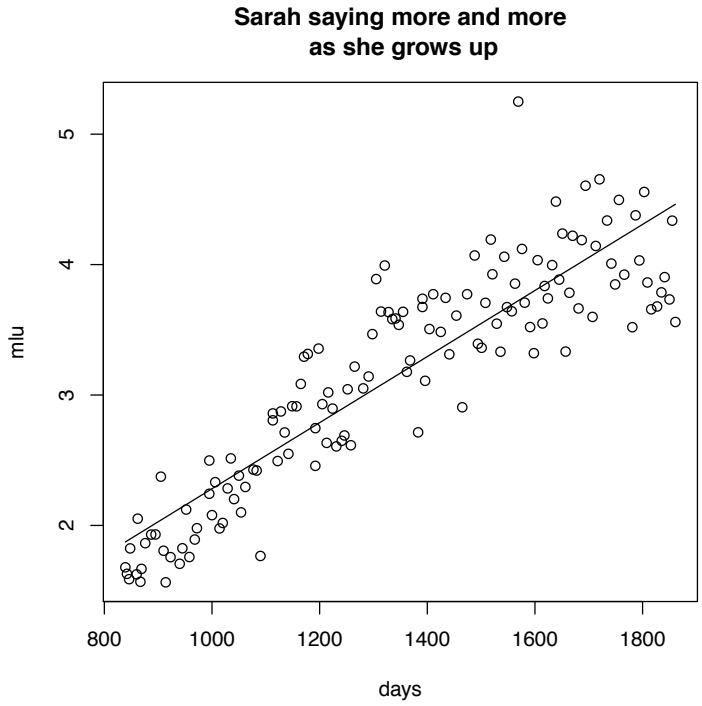
Language Acquisition

In the 1960s Roger Brown and his students recorded three children in the Cambridge, MA area as they were learning their first language, English. One of them was named "Sarah." As part of the CHILDES project, these transcripts are available on-line. When the project started, Sarah was 27 months old.

```
        COU Courtney Investigator, ROG Roger_Brown Investigator
*GLO:   <don't> [/] don't use the words yourself # just ask her .
        pro|her .
*MOT:   c(o)me (h)ere .
*SAR:   xx .
*GLO:   alright .
*SAR:   xx .
*GLO:   yeah .
*GLO:   what's this ?
*SAR:   a nose .
*GLO:   your nose .
*GLO:   an(d) what's that ?
*SAR:   a eye .
*MOT:   0 .
*SAR:   hair .
*MOT:   hair .
*MOT:   where's your teeth ?
*MOT:   oh # what's this ?
*MOT:   what's this in here ?
*MOT:   what's this in here ?
*MOT:   what's that ?
*SAR:   a yy .
*MOT:   what is it ?
*MOT:   what's this ?
*SAR:   yy yy .
*MOT:   what's this ?
*SAR:   tee(th) .
*MOT:   teeth .
```

Is Sarah's mean-length-of-utterance in morphemes (MLU) well-describe by her age?

```
> sarah <- read.table(file = "sarah.txt", header = T)
> head(sarah)

      filename   mlu    age days
1 sarah003.cha 1.679 2;3.19  839
2 sarah004.cha 1.629 2;3.22  842
3 sarah005.cha 1.588 2;3.26  846
4 sarah006.cha 1.824 2;3.28  848
5 sarah007.cha 1.625 2;4.10  860
6 sarah008.cha 2.052 2;4.12  862

> attach(sarah)
> plot(mlu ~ days, main = "Sarah saying more and more\n as she grows up")
> sarahFittedModel <- lm(mlu ~ days)
> lines(days, fitted(sarahFittedModel))
```



The R function `lm` takes a **model formula** of the form

$$\text{response} \sim \text{predictor}_1 + \text{predictor}_2 + \cdots + \text{predictor}_n$$

it gives back a **fitted model object** that can be interrogated using various functions

summary tells t-statistics, standard errors for all estimated parameters. Reports correlation and result of an F-test for the hypothesis that the regression coefficient is zero (same as t-test).

coefficients says what the coefficients actually are

resid calculates the residuals. Should be Normal if a linear model is really right.

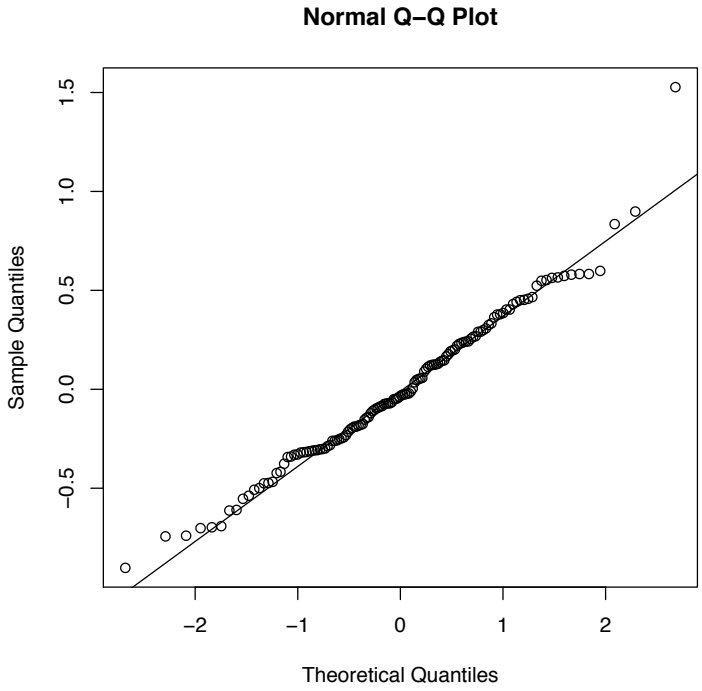
fitted the values that result from stuffing the x values into the equation for the best-fitting line

predict predict new values, possible with confidence intervals (see Dalgaard 5.3)

```
> summary(sarahFittedModel)

Call:
lm(formula = mlu ~ days)

Residuals:
     Min       1Q   Median       3Q      Max
-0.90260 -0.26628 -0.03409  0.24548  1.52713

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.2519622  0.1491562  -1.689   0.0935 .
days         0.0025334  0.0001086  23.334   <2e-16 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.3814 on 134 degrees of freedom
Multiple R-squared: 0.8025,        Adjusted R-squared: 0.801
F-statistic: 544.5 on 1 and 134 DF,  p-value: < 2.2e-16
```

The linear model of Sarah says that her utterances get longer at a rate of about two-hundredths of a morpheme per day, on average.

```
> qqnorm(resid(sarahFittedModel))
> qqline(resid(sarahFittedModel))
```



The 98$^{\text{th}}$ data point comes when Sarah is 4;3.19 (file `sarah101.chA`). In this session, an investigator named Gail Plotkin shows Sarah a picture book and asks her if she has ever seen a falling star. Sarah says "yep." "What does it look like?" Gail asks? Sarah replies "it look like one here 'n one here 'n one here 'n one here 'n one here 'n one there one there one there."...definitely an outlier!

Now You Try

1. In my continuing researches, I have attached a weights of various sizes to the end of a piece of piano wire. Then I measured the length of the wire. My results are collected in table 1

Weight (kg)	Length (cm)
0	439.00
2	439.12
4	439.21
6	439.31
8	439.40
10	439.50

Table 1: Weight vs piano-wire stretchiness

By how much does each additional kilogram cause the wire to stretch? Predict the length under the following loads, if possible: 3 kg, 7 kg, 50 kg