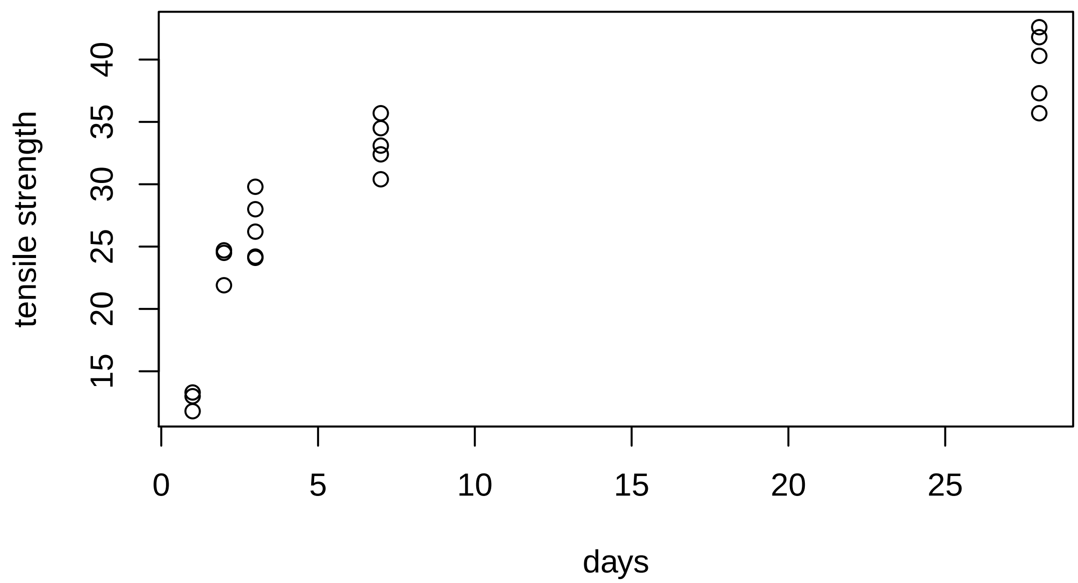
Stat 311: HW on Regression, not due, solutions to be posted before final

Fritz Scholz

The data in `tensile.csv` comes from Problem 10 in Section 15.7 in the text. Read the text there for background information. Download this file (from our class HW site) and load its data into R via
`tensile <- read.csv("tensile.csv",header=T)`
Make sure the file `tensile.csv` resides in the directory from which you start R.

1. Plot the tensile strength against the curing time, labeling the axes appropriately, i.e.,
   `plot(tensile[,1],tensile[,2],xlab="days",ylab="tensile strength")`

   Do the points appear to follow a simple linear regression model?



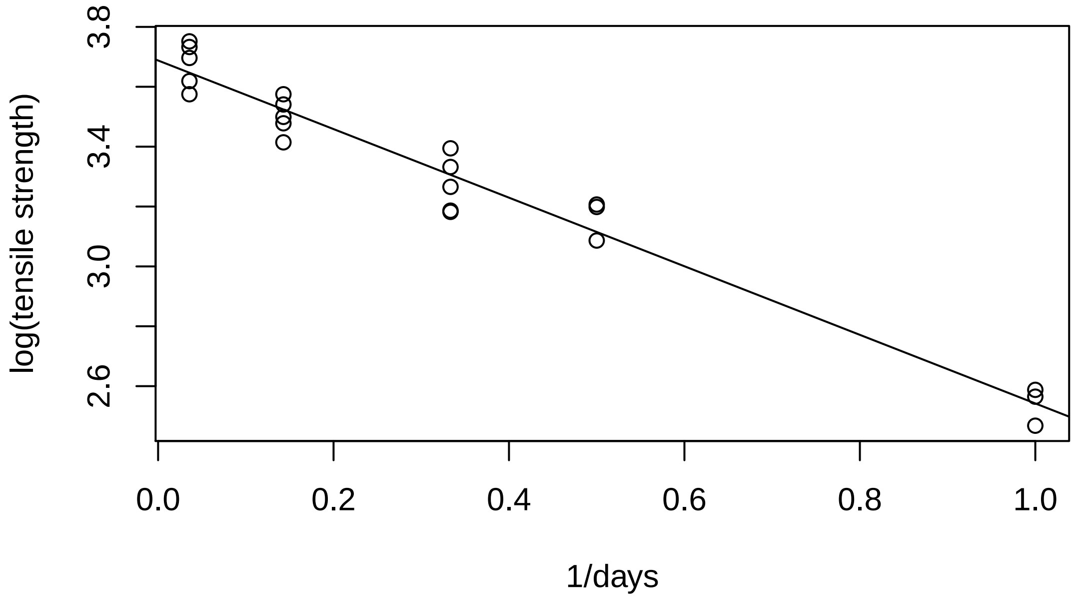
   The plot shows a strongly curved pattern, first a steep increase and then a leveling out.

2. What is n, the number of plotted points?
   $n = 21 < -$`length(tensile[,1])`

3. Make a similar plot of log(tensile strength) against 1/days, labeling the axes correspondingly. Does this plot suggest a simple linear regression model of $y =$ log(tensile strength) in relation to $x =$1/days? For the following let `x <- 1/tensile[,1]` and `y <- log(tensile[,2])`. You can add a fitted regression line to this plot via
   `abline(lsfit(x,y))`
   The commands

```
> plot(x,y,xlab="1/days",ylab="log(tensile strength)")
> abline(lsfit(x,y))
```

produce



The plot looks very linear, i.e., a linear regression model should be adequate.

4. Looking at this last plot, does it suggest that there would be much improvement in tensile strength when using more than 28 days curing time?

5. Find $\sum(x_i - \bar{x})(y_i - \bar{y})$ simply by using `sum((x-mean(x))*(y-mean(y)))` and similarly find $\sum(x_i - \bar{x})^2$, where the summations are over $i = 1, \ldots, n$.

```
> sum((x-mean(x))*(y-mean(y)))
[1] -2.337782
> sum((x-mean(x))^2)
[1] 2.040789
```

6. Find the least squares estimates `beta1.hat` $= \hat{\beta}_1$ and `beta0.hat` $= \hat{\beta}_0$.
Compare the results with `lsfit(x,y)$coef`.

```
> beta1.hat <- sum((x-mean(x))*(y-mean(y)))/sum((x-mean(x))^2)
> beta0.hat <- mean(y)-beta1.hat*mean(x)
> beta1.hat
[1] -1.145528
> beta0.hat
[1] 3.687818
> lsfit(x,y)$coef
Intercept         X
 3.687818 -1.145528
```

`lsfit(x,y)$coef` gives us the same results as are obtained by direct calculation using the provided formulas.

7. Find the vector $\texttt{y.hat} = \texttt{beta0.hat} + \texttt{beta1.hat} * \texttt{x} = (\hat{y}(x_1), \ldots, \hat{y}(x_n))$ of fitted or predicted values for $x_1, \ldots, x_n$, get the vector of residuals $r_i = y_i - \hat{y}(x_i), i = 1, \ldots, n$.
   Compare these with `lsfit(x,y)$resid`. Calculate SS_E and MS_E from these residuals.

```
> y.hat <- beta0.hat+beta1.hat*x
> residuals <- y-y.hat
> y.hat
 [1] 2.542290 2.542290 2.542290 3.115054 3.115054 3.115054 3.305976 3.305976
 [9] 3.305976 3.305976 3.305976 3.524172 3.524172 3.524172 3.524172 3.524172
[17] 3.646907 3.646907 3.646907 3.646907 3.646907
> residuals
 [1]  0.02265927  0.04547395 -0.07419055 -0.02856765  0.08361883  0.09174896
 [7]  0.08853271  0.02622883 -0.12376384 -0.11962305 -0.04021627 -0.04601314
[13] -0.10972896  0.01678776 -0.02463829  0.05097912  0.08598959  0.10494750
[19]  0.04944472 -0.07175606 -0.02791343
> lsfit(x,y)$resid
 [1]  0.02265927  0.04547395 -0.07419055 -0.02856765  0.08361883  0.09174896
 [7]  0.08853271  0.02622883 -0.12376384 -0.11962305 -0.04021627 -0.04601314
[13] -0.10972896  0.01678776 -0.02463829  0.05097912  0.08598959  0.10494750
[19]  0.04944472 -0.07175606 -0.02791343
# with exactly the same residuals
> SS.E <- sum(residuals^2)
> MS.E <- SS.E/(21-2)
> SS.E
[1] 0.1085086
> MS.E
[1] 0.00571098
```

8. Get a 95% confidence interval for the slope parameter β_1 in this transformed variables regression situation. Should the hypothesis $H_0 : \beta_1 = 0$ be rejected at level $\alpha = 0.05$?

```
> qt(0.975,21-2)
[1] 2.093024
> t.xx <- sum((x-mean(x))^2)
> beta1.hat-qt(0.975,21-2)*sqrt(MS.E/t.xx)
[1] -1.256250
> beta1.hat+qt(0.975,21-2)*sqrt(MS.E/t.xx)
[1] -1.034807
# with 95% confidence interval (-1.256250, -1.034807).
# It does not contain beta.1 = 0, thus reject that hypothesis at alpha=0.05.
```

9. Get a 95% confidence interval for the mean $\mu_y(x = 1/28)$.

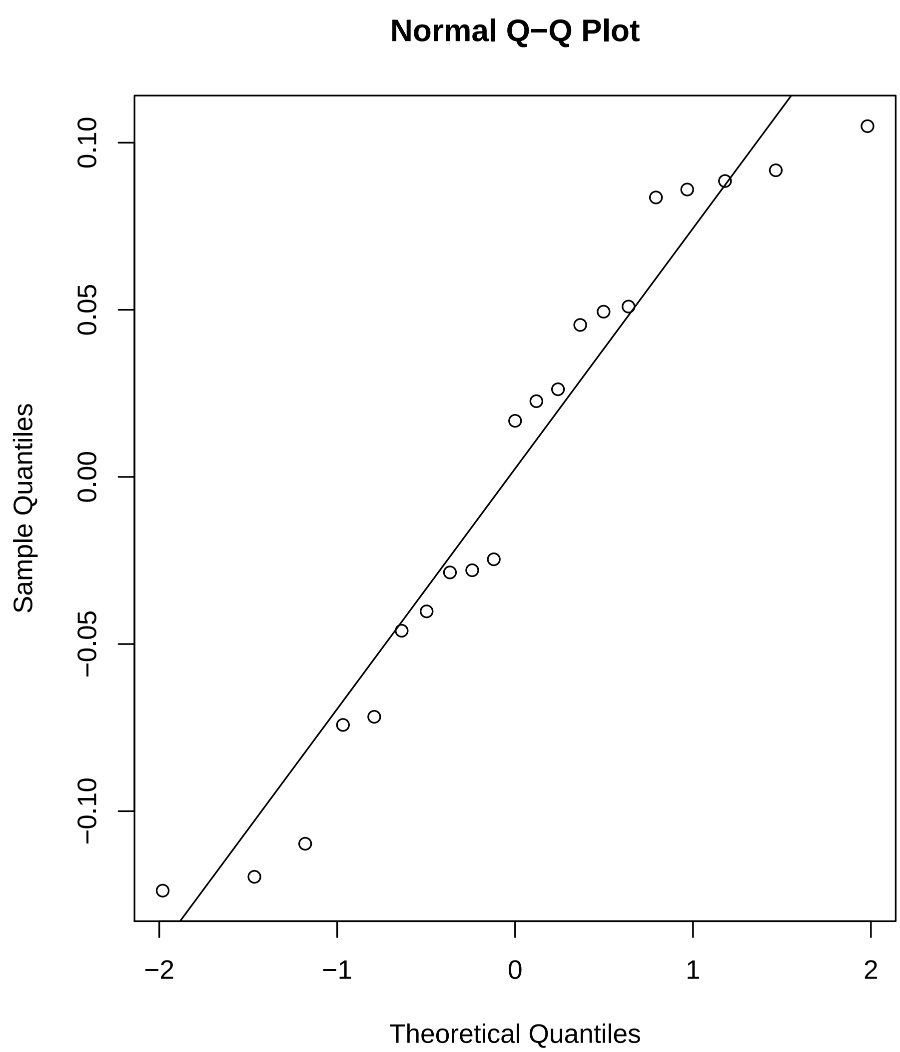
```
> y.hat.28 <- beta0.hat+beta1.hat*(1/28)
> y.hat.28
[1] 3.646907
> y.hat.28-qt(0.975,21-2)*sqrt(MS.E*(1/21+(1/28-mean(x))^2/t.xx))
[1] 3.598969
```

```
> y.hat.28+qt(0.975,21-2)*sqrt(MS.E*(1/21+(1/28-mean(x))^2/t.xx))
[1] 3.694844
# with confidence interval (3.598969, 3.694844) for the mean log(tensile strength)
# after 28 days of curing.
```

10. Transform back the last interval into a corresponding one for tensile strength at 28 days.

```
> exp(3.598969)
[1] 36.56052
> exp(3.694844)
[1] 40.23929
# giving us as interval (36.56052, 40.23929) for what?
```

We should not interpret $\exp(\mu_y(x))$ as the mean of the tensile strength since $\exp(E(\log(T))) \neq E(T)$ or $E(\log(T)) \neq \log(E(T))$ where T represents the tensile strength at x.



However, the normal probability plot of the residuals suggests that the normality assumption for the error term in the simple linear regression model is justified. Thus the mean of the log-tensile strength at each x can also be viewed as the median at each x. Transforming such log-tensile strengths back does not affect the median character, since 50% will be above and below the median at any x, before or after back-transformation via exp(). Thus the above interval can be viewed as a 95% confidence interval for the median tensile strength after 28 days.