Solutions: Meat Processing and pH

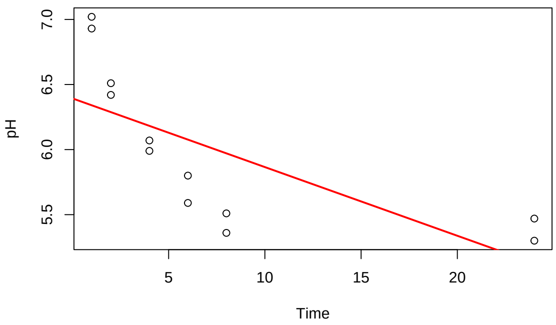
Begin by importing the `meatph.csv` data file into R.

1. Consider a simple linear model to use `Time` to predict `pH`.

   a. Make a scatterplot of `pH` vs. `Time`. Describe any patterns observed.

   R code for parts (a) and (b) is below. The is a clear negative association between the variables, with `pH` decreasing as `Time` increases. The pH decreases rapidly for the first 8 hours after slaughter, but then changes very little between 8 hours and 24 hours. The relationship, however, is clearly non-linear, and there are two very unusual points at about 24 hours after slaughter. Those unusual points fall well outside the pattern of the rest of the data. The line fit in part (b) does a very poor job of reflecting the patterns in the data.

```
plot(pH ~ Time, data=meatph)
model <- lm(pH ~ Time, data=meatph)
abline(model, col="red", lwd=2)
```



   b. Find the least squares regression line for predicting `pH` from `Time`. Add a plot of the line to your scatterplot from question 1.

   See above for R code and plot. From the output below, the equation for the regression line is

$$\widehat{pH} = 6.393 - 0.05277 \times Time.$$

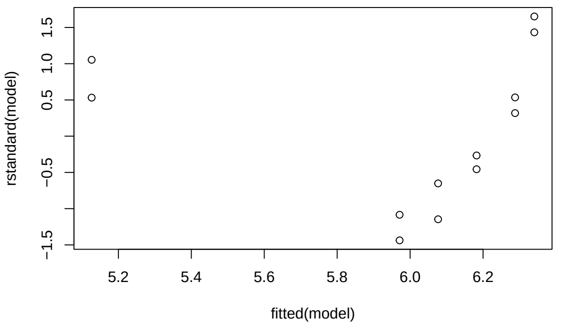
```
model
```

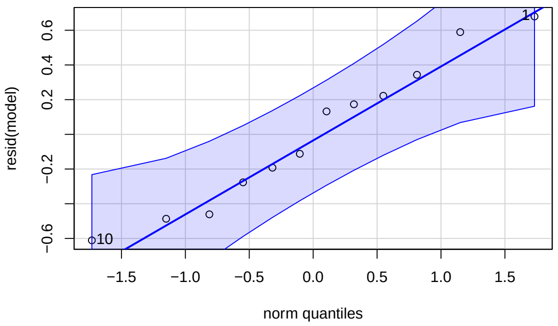
```
##
## Call:
## lm(formula = pH ~ Time, data = meatph)
##
## Coefficients:
## (Intercept)         Time
##     6.39331     -0.05277
```

   c. Do any of the conditions for the simple linear model appear to be violated? If so, which conditions? Explain your answer, and include any graphs that might be helpful in supporting your conclusion.

```
plot(rstandard(model) ~ fitted(model))
library(car)
qqPlot(resid(model))
```

```
## [1]  1 10
```





Based on both the scatterplot in part (a) and the residuals vs. fitted values plot above, the pattern in the data is clearly non-linear, so a linear model is not appropriate. The residuals do appear to be reasonably close to normally distributed. However, there are also the two unusual points noted earlier, which will be discussed in more detail in part (d) below.

d. Do there appear to be any unusual points in the linear model? If so, are these points outliers? Do they have high leverage? Are they influential points? Clearly explain your answers.

   There are two very unusual points recorded 24 hours after slaughter which fall well outside the pattern of the data.

   - These points are NOT outliers in the regression model. As indicated by the plot of the standardized residuals above, the residuals for these points are NOT unusually large.
   - The points do have extremely high leverage due to their very extreme x values.
   - The points are highly influential. We see in the plot in part (a) that the regression line does not at all reflect the pattern of the rest of the data. Without the two unusual points, the regression line would have a much steeper negative slope to follow the pattern of the points from 0 to 8 hours after slaughter. The line is pulled upward substantially toward the two unusual points.

2. Consider the two carcasses for which the pH was measured 24 hours after slaughter. Can you think of any strong justification(s) for removing these two points from the model? (Hint: There are at least two distinct good reasons!)

   Two reasons:

   - The whole purpose of the study was to estimate the time needed for the pH to drop to 6.0. That has clearly already occurred by about 4-6 hours after slaughter, so what happens after 24 hours is not relevant to the purpose of the study.
   - As discussed in the notes/lecture, it is generally reasonable to remove points that are very extreme in the *predictor* variable, which is the case with these points. We can build a model that works well for 0 to 8 hours after slaughter and avoid using the model for longer time periods than that.

3. Create a new dataset named `meatph2` which removes the two carcasses for which the pH was measured 24 hours after slaughter. Repeat the four steps from question 1 with this new dataset with these two points removed.

   We create the dataset as follows:

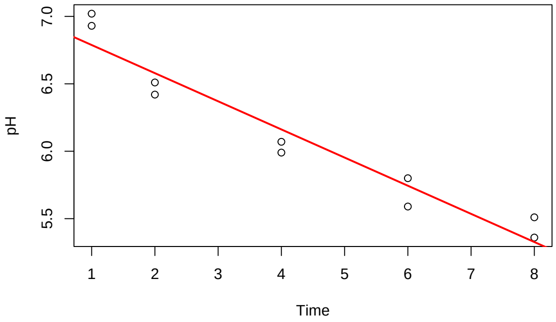
```r
meatph2 <- subset(meatph, Time < 24)
```

a. Make a scatterplot of `pH` vs. `Time`. Describe any patterns observed.

   R code for parts (a) and (b) is below. The is an even clearer negative association than before between the variables, with `pH` decreasing as `Time` increases. The relationship, however, is still somewhat non-linear. It decrease more rapdly at the start and then starts to flatten out a bit.

The line fit in part (b) reflects the general overall decreasing trend, but misses the "curvature" of the pattern. It under-predicts the pH at the start ($t = 0$) and end ($t = 8$) but tends to over-predict in between.

```
plot(pH ~ Time, data=meatph2)
model2 <- lm(pH ~ Time, data=meatph2)
abline(model2, col="red", lwd=2)
```



b. Find the least squares regression line for predicting `pH` from `Time`. Add a plot of the line to your scatterplot from question 1.

See above for R code and plot. From the output below, the equation for the regression line is

$$\widehat{pH} = 6.9965 - 0.2087 \times Time.$$

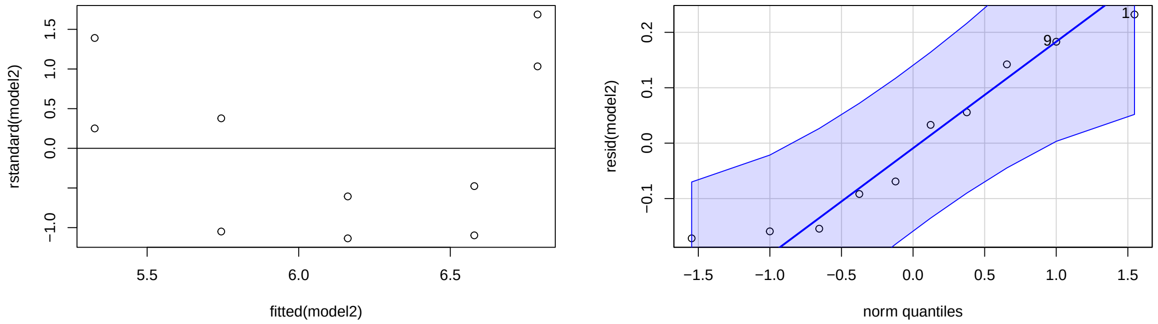
```
model2
```

```
##
## Call:
## lm(formula = pH ~ Time, data = meatph2)
##
## Coefficients:
## (Intercept)         Time
##      6.9965      -0.2087
```

c. Do any of the conditions for the simple linear model appear to be violated? If so, which conditions? Explain your answer, and include any graphs that might be helpful in supporting your conclusion.

```
plot(rstandard(model2) ~ fitted(model2))
abline(h=0)
qqPlot(resid(model2))
```

```
## [1] 1 9
```



Based on both the scatterplot in part (a) and the residuals vs. fitted values plot above, the pattern in the data is non-linear, violating the linearity condition. The residuals do appear to be reasonably close to normally distributed. There no longer appear to be any outliers or high leverage points. There might be a hint of heterscedasticity, with more veriability in the residuals for smaller fitted values and less variability for larger fitted values. The key problem is the curvature in the residual plot.

d. Do there appear to be any unusual points in the linear model? If so, are these points outliers? Do they have high leverage? Are they influential points? Clearly explain your answers.

   No, there are no unusual points.

4. Note that pH is a logarithmic scale (similar to other commonly used measurements like decibels for sound or the Richter scale for earthquakes). When trying to model a logarithmic scale in linear regression, it is sometimes helpful to also take a logarithm of the predictor variable so that both variables are on logarithmic scales. Create a new variable named `logTime` in the dataset by running the command below. (Note that in R, `log(x)` is the natural logarithm (usually written ln(x)) with base e, while log10(x) is the logarithm with base 10.)

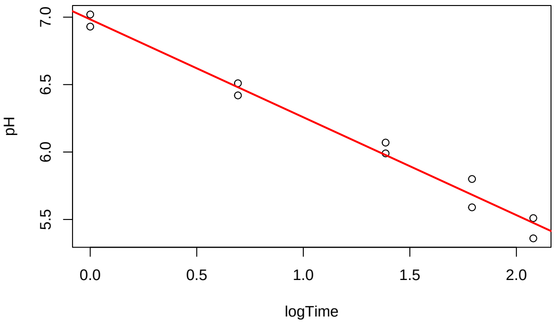
```
meatph2$logTime = log(meatph2$Time)
```

Repeat the four steps from question 1, but use `logTime` as the explanatory variable rather than `Time` (since you are using `meatph2`, you will also be omitting the two 24 hour points). For parts (a) and (b), plot `pH` vs. `logTime` in the scatterplot.

a. Make a scatterplot of `pH` vs. `logTime`. Describe any patterns observed.

R code for parts (a) and (b) is below. There is a very strong, negative, linear association between the variables, with `pH` decreasing linearly as `logTime` increases. The line fit in part (b) reflects the pattern of the data very well. The transformation appears to have addressed the curvature observed in the previous model.

```
plot(pH ~ logTime, data=meatph2)
model3 <- lm(pH ~ logTime, data=meatph2)
abline(model3, col="red", lwd=2)
```



b. Find the least squares regression line for predicting `pH` from `logTime`. Add a plot of the line to your scatterplot from question 1.

See above for R code and plot. From the output below, the equation for the regression line is

$$\widehat{pH} = 6.9836 - 0.7257 \times logTime.$$

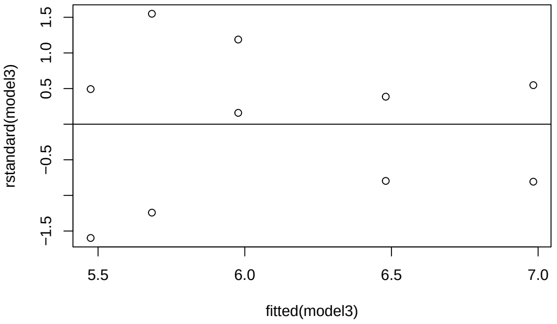
```
model3
```

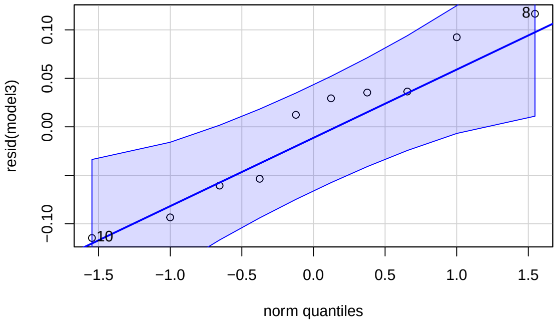
```
##
## Call:
## lm(formula = pH ~ logTime, data = meatph2)
##
## Coefficients:
## (Intercept)      logTime
##      6.9836      -0.7257
```

c. Do any of the conditions for the simple linear model appear to be violated? If so, which conditions? Explain your answer, and include any graphs that might be helpful in supporting your conclusion.

```
plot(rstandard(model3) ~ fitted(model3))
abline(h=0)
qqPlot(resid(model3))
```

```
## [1]  8 10
```





There are no clear patterns in the residuals and the relationship does appear to be linear. The residuals do appear to be reasonably close to normally distributed. There are no outliers or high leverage points. There might be a hint of heterscedasticity, with more veriability in the residuals for smaller fitted values and less variability for larger fitted values, but it is not particularly strong. Overall, the conditions appear to be met pretty well by this model.

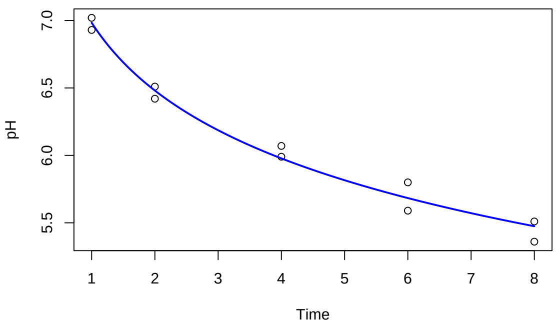
d. Do there appear to be any unusual points in the linear model? If so, are these points outliers? Do they have high leverage? Are they influential points? Clearly explain your answers.

   No, there are no unusual points.

5. The model found in question 4 has the form $\widehat{pH} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \ln(Time)$. In R, if you stored the results of the `lm()` command as an object named `model`, you can get the coefficients for the model using the commands `coef(model)[1]` for the intercept $\hat{\beta}_0$ and `coef(model)[2]` for the slope $\hat{\beta}_1$. (If you named the object something other than `model`, you will need to replace `model` in those commands with the name of your object.)

   Use those coefficients to create a plot of `pH` vs. `Time` showing the transformed model:

```
plot(pH ~ Time, data = meatph2)
curve(coef(model3)[1] + coef(model3)[2]*log(x), add=T, col="blue", lwd=2)
```



6. Which of the three models that you produced do you think is the "best" model for predicting pH as a function of time? Explain. Use that model to estimate the amount of time it takes on average after slaughter for the pH to drop to 6.0.

   The final model produced in question 4 is the best model. It ignores the unusual points at 24 hours which were not relevant for the quesiton of interest and which had high influence. It also addresses the problems with lack of linearity in the relationship. It very accurately reflects the pattern of the relationship as shown in the plot in question 5.

   Based on the model, we have

$$\widehat{pH} = 6.9836 - 0.7257 \times logTime,$$

   so setting with pH $= 6$, we have

$$6 = 6.9836 - 0.7257 \times \ln(Time).$$

   Solving this equation for $Time$ yields

$$-0.9836 = -0.7257 \times \ln(Time) \implies \frac{-0.9836}{-0.7257} = \ln(Time) \implies 1.355 = \ln(Time) \implies Time = e^{1.355} \approx 3.9$$

   Thus the model predicts that it will typically take about 3.9 hours after slaughter for the pH to drop to 6.0.