# Problem Set 2

*Math 430, Winter 2017*

## 2.2

First, we must load the data found in `indicators.txt` on the textbook's webpage.

```
indicators <- read.table("http://www.stat.tamu.edu/~sheather/book/docs/datasets/indicators.txt",
                         header = TRUE)
```

Next, we must fit the model $Y = \beta_0 + \beta_1 x + e$.

```
mod2.2 <- lm(PriceChange ~ LoanPaymentsOverdue, data = indicators)
```

### (a)

```
confint(mod2.2)
```

```
##                         2.5 %     97.5 %
## (Intercept)         -2.532112 11.5611000
## LoanPaymentsOverdue -4.163454 -0.3335853
```

We are 95% confident that for a 1% increase in the mortgage loans 30 days or more overdue in latest quarter, we expect between a 0.33% and 4.16% decrease in the average price from July 2006 to July 2007. Since the confidence interval does not contain 0, there is evidence of a significant negative linear association at the $alpha = 0.05$ level.

### (b)

We are asked for a 95% confidence interval for $E(Y|X = 4)$, which we can find easily using the `predict` function.

```
predict(mod2.2, newdata = data.frame(LoanPaymentsOverdue = 4), interval = "confidence")
```

```
##         fit       lwr       upr
## 1 -4.479585 -6.648849 -2.310322
```

The confidence interval for $E(Y|X = 4)$ does not contain 0, so 0% is not a feasible value.

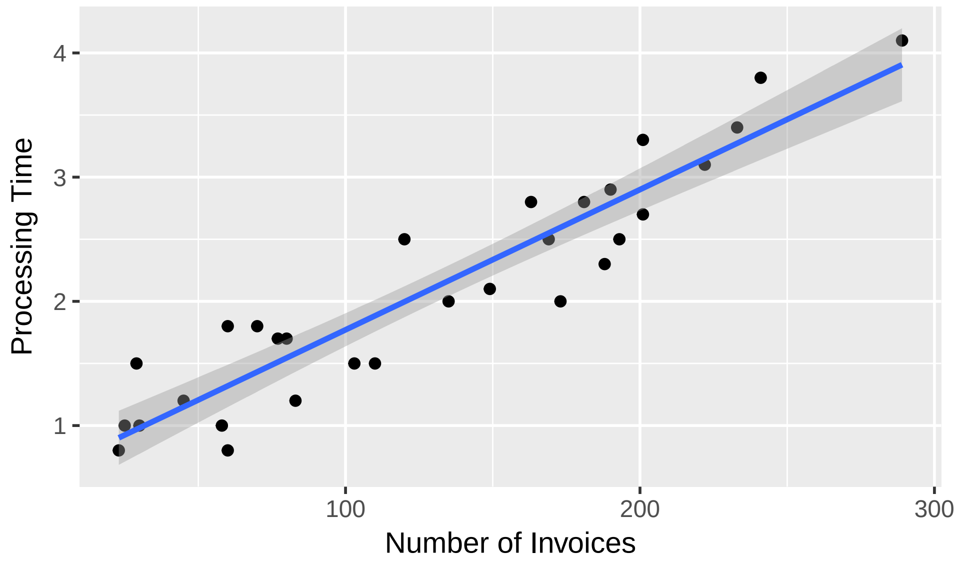
---

## 2.3

```
invoices <- read.table("http://www.stat.tamu.edu/~sheather/book/docs/datasets/invoices.txt",
                       header = TRUE)
```

We can easily replicate the output displayed in the textbook.

```
library(ggplot2)
ggplot(data = invoices, aes(x = Invoices, y = Time)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Number of Invoices", y = "Processing Time")
```



```r
mod2.3 <- lm(Time ~ Invoices, data = invoices)
summary(mod2.3)
```

```
##
## Call:
## lm(formula = Time ~ Invoices, data = invoices)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59516 -0.27851  0.03485  0.19346  0.53083
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.6417099  0.1222707   5.248 1.41e-05 ***
## Invoices    0.0112916  0.0008184  13.797 5.17e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3298 on 28 degrees of freedom
## Multiple R-squared:  0.8718, Adjusted R-squared:  0.8672
## F-statistic: 190.4 on 1 and 28 DF,  p-value: 5.175e-14
```

### (a)

We can calculate confidence interval for regression coefficients using the `confint` function.

```r
confint(mod2.3)
```

```
##                   2.5 %     97.5 %
## (Intercept) 0.391249620 0.89217014
## Invoices    0.009615224 0.01296806
```

So a 95% confidence interval for the y-intercept, $\beta_0$, is (0.39, 0.89).

Alternatively, we could calculate the interval by hand (which you should understand how to do!).

$$\widehat{\beta_0} \pm t^*_{\alpha/2,n-2} se(\widehat{\beta}_0) = 0.6417099 \pm 2.048407(0.1222707) = (0.3912497, 0.8921701)$$

where we find $t^*_{\alpha/2,n-2}$ using the `qt` function below.

```
qt(.975, 28)
```

```
## [1] 2.048407
```

The intervals agree up to the expected rounding error.

### (b)

We wish to run a test of the hypotheses

$$H_0 : \beta_1 = 0.01 \qquad \text{vs.} \qquad H_A : \beta_1 \neq 0.01$$

First, we must calculate the test statistic.

$$T = \frac{\widehat{\beta}_1 - \beta_1^0}{se(\widehat{\beta}_1)} = \frac{0.0112916 - 0.01}{0.0008184} \approx 1.578$$

Next, we compare our test statistic to it's sampling distribution assuming that $H_0$ is true. A common way to do this is to calculate the p-value (i.e the probability of observing a $T$ that is at least as far from 0 in magnitude as what we have observed).

```
2 * (1 - pt(1.578, df = 28))
```

```
## [1] 0.125798
```

Based on a test statistic of $T = 1.578$ and associated p-value of 0.0629, there is no evidence that the slope is not 0.01. That is, we find no evidence of a difference from the best practice benchmark.

### (c)

To find a point estimate and a 95% prediction interval for the time taken to process 130 invoices we can use the `predict` function (but you should be able to do this by hand, given summary statistics):

```
predict(mod2.3, newdata = data.frame(Invoices = 130), interval = "prediction")
```

```
##        fit      lwr    upr
## 1 2.109624 1.422947 2.7963
```

The 95% prediction interval for the time taken to process 130 invoices is 1.4 to 2.8 hours.

---

## 3.3

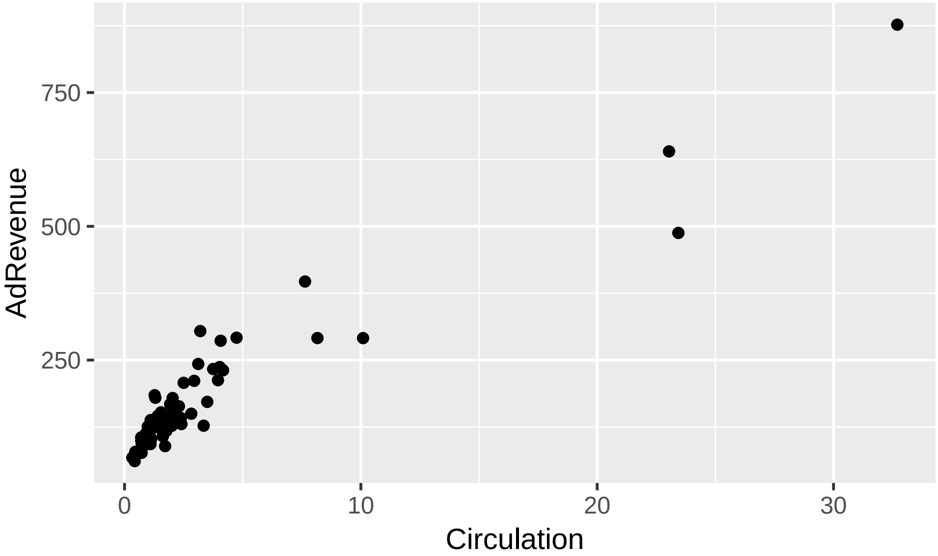
```
ad_revenue <- read.csv("http://www.stat.tamu.edu/~sheather/book/docs/datasets/AdRevenue.csv")
```

**(a)**

Multiple answers are accepted here as long as you provided justification for your model. We had not yet covered transformations, so I did not expect an elaborate search. Rather, I expected that you would plot the data and go from there.

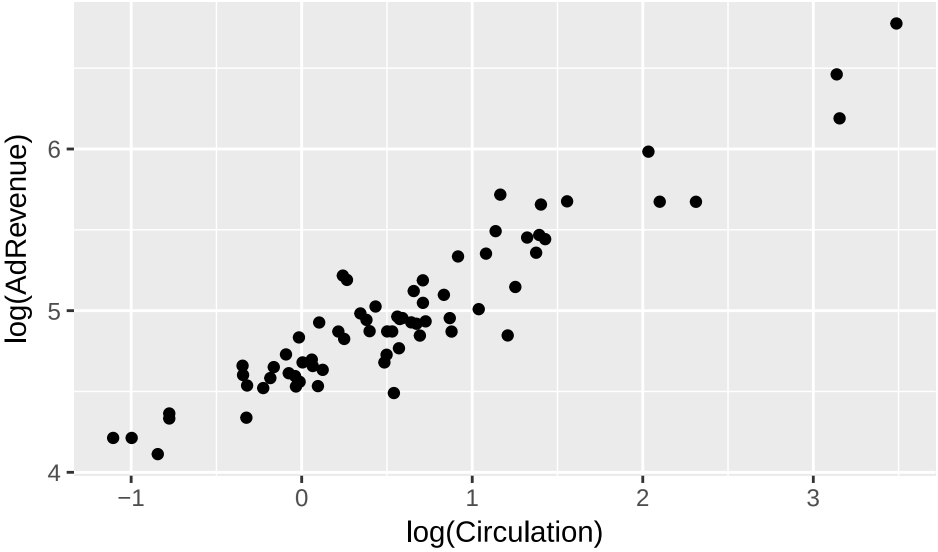
The plot of the raw data (shown below) appears to exhibit curvature, indicating a nonlinear relationship between circulation and ad revenue.

```
ggplot(data = ad_revenue, aes(x = Circulation, y = AdRevenue)) +
  geom_point()
```



In an effort to find a linear relationship you may have tried to log-transform the variables, which appears to be relatively successful at linearizing the relationship.

```
ggplot(data = ad_revenue, aes(x = log(Circulation), y = log(AdRevenue))) +
  geom_point()
```



The log transformed model is fit below

```
mod3.3_tform <- lm(log(AdRevenue) ~ log(Circulation), data = ad_revenue)
```

I also fit the untransformed model, just for comparison.

```
mod3.3_raw<- lm(AdRevenue ~ Circulation, data = ad_revenue)
```

### (b)

We can use the `predict` function to obtain the prediction intervals.

```
tformed_preds <- predict(mod3.3_tform, newdata = data.frame(Circulation = c(0.5, 20)), interval = "pred
tformed_preds
```

```
##        fit      lwr      upr
## 1 4.308227 3.947855 4.668600
## 2 6.258752 5.885815 6.631689
```

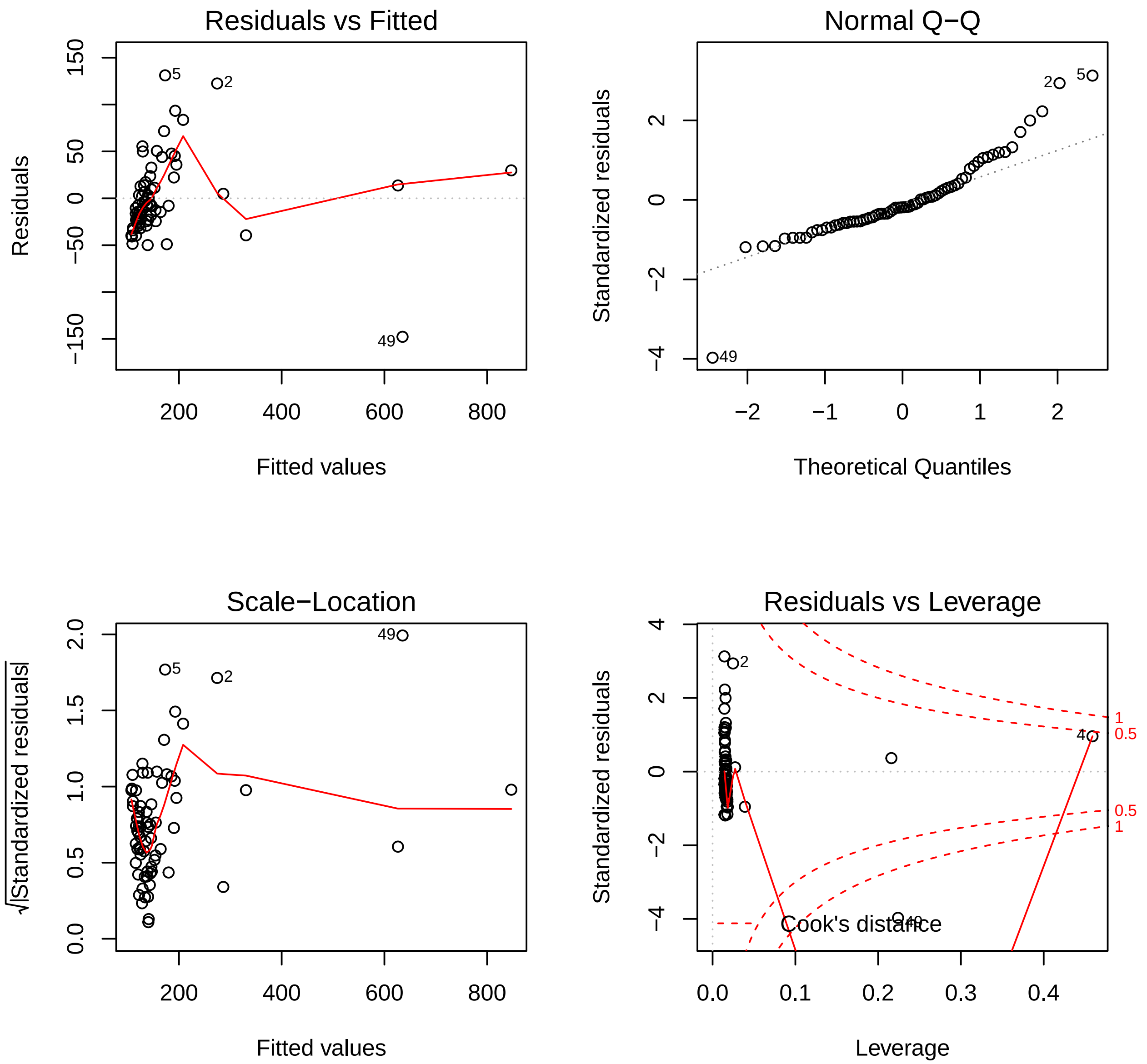
We need to back-transform the intervals to the original scale for interpretability:

```
exp(tformed_preds)
```

```
##         fit       lwr      upr
## 1  74.30864  51.82406 106.5485
## 2 522.56626 359.89585 758.7626
```

### (c)

If you you did not transform your model, then you would obtain the below set of residual plots

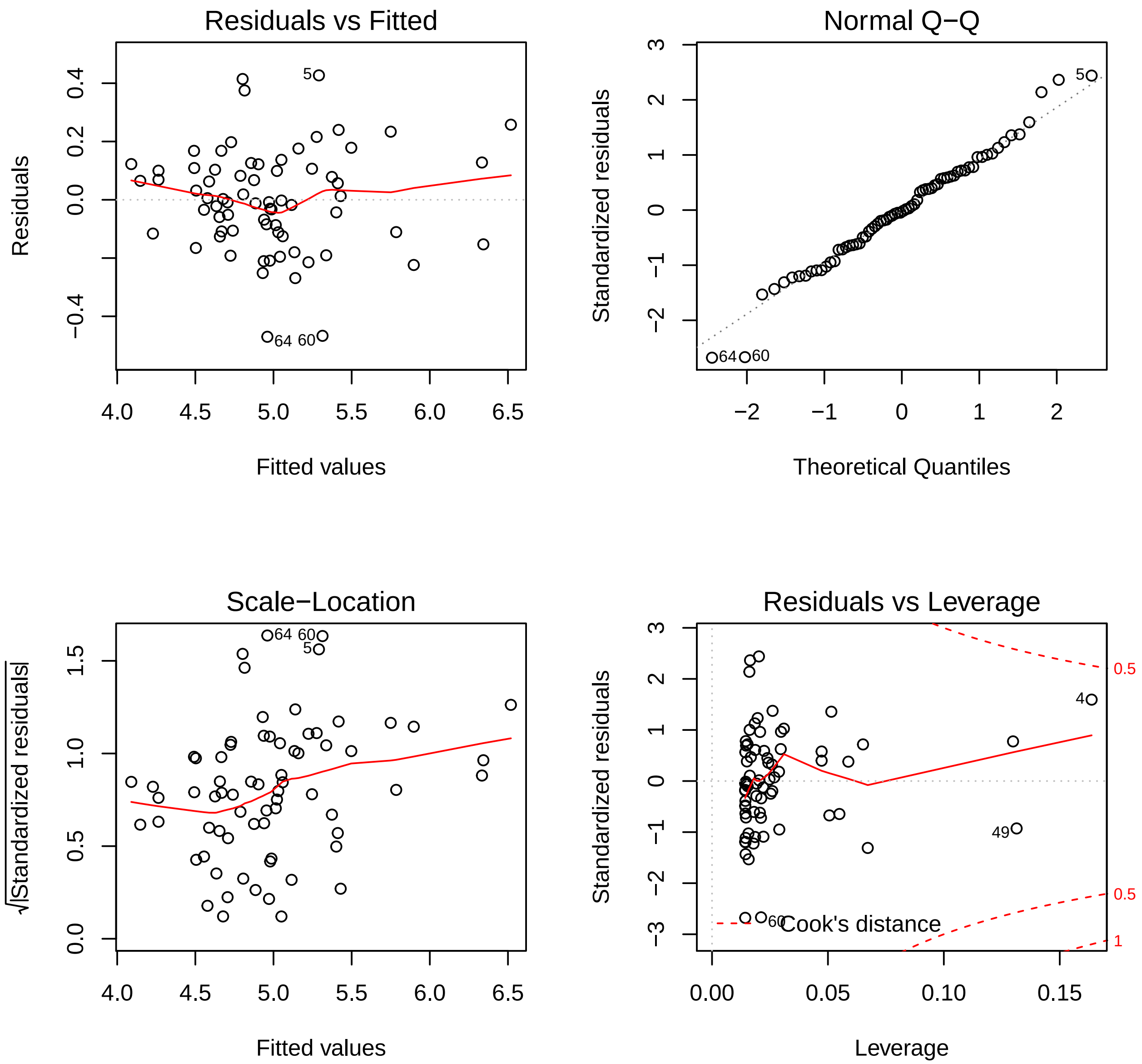


```
## null device
##           1
```

Weaknesses of this model include:

- The residuals do not appear to be randomly scattered about 0 in the plot of the residuals vs. circulation. There appear to be far more negative residuals at lower circulations, which is indicative of a nonlinear relationship. (We saw this nonlinear relationship above!)
- The scale-location plot seems to indicate that the variance of the residuals is not constant, especially for the residuals associated with fitted values below 200.
- The normal distribution does not well-approximate the residual distribution. The residuals appear to be heavier-tailed than the normal distribution.
- There are potentially influence observations.

If you log-transformed both variables in your model, then you would obtain the below set of residual plots



```
## null device
##           1
```

Weaknesses of this model include:

- The residuals may not have constant variance, as seen by the upward trend on the scale-location plot.

## 3.4

```
glakes <- read.table("http://www.stat.tamu.edu/~sheather/book/docs/datasets/glakes.txt", header = TRUE)
```

### (a)

Weaknesses of this model include:

- The residuals appear to have nonconstant variance, as seen by the upward trend on the scale-location plot and the "megaphone pattern" in the plot of the standardized residuals.

### (b)

Since we identified nonconstant error variance in part (a), we know that prediction intervals are suspect. Notice first that we see increasing variance about the line as Tonnage increases. Further, note that the average value of tonnage is $\overline{x} = 3416.5$ and that we wish to predict at $x = 10,000$. The estimation process used to estimate the residual standard deviation is an issue here, as it is summarizing the variability across the entire range of $x$ using a single value; consequently, it will likely be too large in the region with little variability and too small in regions with higher variability. Since $x = 10,000$ is far to the right and in a region of high variability, we expect the prediction interval to be too small.