STAT 425 Exam 1 Study Problem Solutions

Exam problems are generally be shorter than homework problems and may involve short answer conceptual questions, quick calculations and R code interpretation or debugging. It is not a multiple choice exam, although some multiple choice questions are possible.

The sample problems below are to help you test yourself and practice solving. Problems on the exam will generally have fewer parts to them than the ones below. Do not expect the actual exam problems to be exactly like this set in terms of range of coverage or length. Work on these various problems as a way to solidify your understanding.

1. Twenty chicks (baby chickens) were randomly assigned to receive one of two diets, A or B, with 10 in each group. Consider the model

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, 3, \ldots, 20.$$

Here y_i denotes the 14-day weight gain for the ith chick, and

$$x_i = \begin{cases} -1, & \text{if chick } i \text{ receives Diet A;} \\ 1, & \text{if chick } i \text{ receives Diet B.} \end{cases}$$

The data are arranged so that Chick numbers 1 - 10 received Diet A and Chick numbers 11-20 received Diet B.

a) Calculate $\bar{x}$ and S_{xx} for this design.

$$\bar{x} = \frac{1}{20}\left(\sum_{i=1}^{10}(-1) + \sum_{i=11}^{20}(1)\right) = \frac{-10+10}{20} = 0$$

$$S_{xx} = \sum_{i=1}^{20}(x_i - \bar{x})^2 = \sum_{i=1}^{10}(-1-0)^2 + \sum_{i=11}^{20}(1-0)^2 = 20$$

b) Show that $\hat{\beta}_1$ the least squares estimate of β_1 equals $\frac{1}{2}(\bar{y}_B - \bar{y}_A)$, where $\bar{y}_A$ and $\bar{y}_B$ are the sample means for weight gain on Diet A and Diet B, respectively.

$$S_{xy} = \sum_{i=1}^{20}(x_i - \bar{x})(y_i - \bar{y}) = -\sum_{i=1}^{10}(y_i - \bar{y}) + \sum_{i=11}^{20}(y_i - \bar{y}) = -10\bar{y}_A + 10\bar{y} + 10\bar{y}_B - 10\bar{y}$$
$$= 10(\bar{y}_B - \bar{y}_A)$$

$$\Rightarrow \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{1}{2}(\bar{y}_B - \bar{y}_A)$$

c) Suppose $\bar{y}_A = 101.2$, $\bar{y}_B = 123.7$ and $\sum_{i=1}^{20}(y_i - \hat{y}_i)^2 = 49.0$. Calculate the value of the t-statistic for testing the null hypothesis that $\beta_1 = 0$.

$$\hat{\beta}_1 = \frac{1}{2}(123.7 - 101.2) = 11.25$$

$$se(\hat{\beta}_1) = \sqrt{\frac{RSS/(20-2)}{S_{xx}}} = \sqrt{\frac{49/18}{20}} = 0.369$$

$$\Rightarrow t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{11.25}{0.369} = 30.5$$

2. We can rewrite the model from Problem 1 in matrix form as $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$, where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{19} \\ y_{20} \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & -1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_{19} \\ e_{20} \end{pmatrix}$$

a) Show that for this design the columns of $\mathbf{X}$ are orthogonal to each other.

Calculate the inner product between the two columns of $\mathbf{X} = (\mathbf{X}_{\cdot 1}\ \mathbf{X}_{\cdot 2})$:

$$\mathbf{X}_{\cdot 1}^T \mathbf{X}_{\cdot 2} = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ -1 \\ \vdots \\ 1 \\ 1 \end{pmatrix} = -10 + 10 = 0$$

The inner product is zero, so the two vectors are orthogonal.

b) Show that for this design $cov(\hat{\beta}_0, \hat{\beta}_1) = 0$.

First let's get the form of the covariance matrix:

$$cov\left(\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}\right) = \sigma^2 \left(\mathbf{X}^\mathbf{T}\mathbf{X}\right)^{-1} = \sigma^2 \begin{pmatrix} 20 & 0 \\ 0 & 20 \end{pmatrix}^{-1} = \sigma^2 \begin{pmatrix} \frac{1}{20} & 0 \\ 0 & \frac{1}{20} \end{pmatrix}$$

The $[1, 2]$ element of the matrix $= 0 = cov(\hat{\beta}_0, \hat{\beta}_1)$.

c) Find the leverage of the first observation. Recall that $h_i = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_i$.

$$h_i = \begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \frac{1}{20} & 0 \\ 0 & \frac{1}{20} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} \frac{1}{20} & -\frac{1}{20} \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \frac{2}{20} = 0.10$$

3. Consider a model of the form $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$, where $\mathbf{X}$ is an $n \times p$ full rank matrix (its columns are linearly independent), $\mathbf{y}$ and $\mathbf{e}$ are $n \times 1$, and β is $p \times 1$. The least squares estimator $\hat{\beta}$ solves the matrix equation

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{0}.$$

Let $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$. Show or explain why each of the following equations holds, using the least squares equation as a starting point:

a) $\hat{\mathbf{y}}^T(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{0}$

Using the LS equation we have:

$$\hat{\mathbf{y}}^\mathbf{T}(\mathbf{y} - \hat{\mathbf{y}}) = \hat{\beta}^\mathbf{T}\mathbf{X}^\mathbf{T}(\mathbf{y} - \mathbf{X}\hat{\beta}) = \hat{\beta}^\mathbf{T}\mathbf{0} = 0$$

b) $\hat{\mathbf{y}}^T\mathbf{y} = \hat{\mathbf{y}}^T\hat{\mathbf{y}}$

From Part a) we have:

$$0 = \hat{\mathbf{y}}^\mathbf{T}(\mathbf{y} - \hat{\mathbf{y}}) = \hat{\mathbf{y}}^\mathbf{T}\mathbf{y} - \hat{\mathbf{y}}^\mathbf{T}\hat{\mathbf{y}}$$

c) $(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{y}^\mathbf{T}\mathbf{y} - \hat{\mathbf{y}}^\mathbf{T}\hat{\mathbf{y}}$

$$\begin{aligned}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) &= (\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}}) \\ &= \mathbf{y}^\mathbf{T}\mathbf{y} - \mathbf{y}^\mathbf{T}\hat{\mathbf{y}} - \hat{\mathbf{y}}^\mathbf{T}\mathbf{y} + \hat{\mathbf{y}}^\mathbf{T}\hat{\mathbf{y}} \\ &= \mathbf{y}^\mathbf{T}\mathbf{y} - 2\hat{\mathbf{y}}^\mathbf{T}\hat{\mathbf{y}} + \hat{\mathbf{y}}^\mathbf{T}\hat{\mathbf{y}} \\ &= \mathbf{y}^\mathbf{T}\mathbf{y} - \hat{\mathbf{y}}^\mathbf{T}\hat{\mathbf{y}}\end{aligned}$$

d) $RSS = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \mathbf{y}^\mathbf{T}\mathbf{y} - \hat{\beta}^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{X}\hat{\beta}$

Using Part c):

$$\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = (\mathbf{y} - \hat{\mathbf{y}})^T(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{y}^\mathbf{T}\mathbf{y} - \hat{\mathbf{y}}^\mathbf{T}\hat{\mathbf{y}} = \mathbf{y}^\mathbf{T}\mathbf{y} - \hat{\beta}^\mathbf{T}\mathbf{X}^\mathbf{T}\mathbf{X}\hat{\beta}$$

e) Which, if any, of equations a), b), c), or d) says that the vector of residuals and vector of fitted values are orthogonal to each other?

Part a) shows this because $\hat{\mathbf{y}}$ is the vector of fitted values and $\mathbf{y} - \hat{\mathbf{y}}$ is the vector of residuals.

4. Consider a model of the form $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$, where $\mathbf{X}$ is an $n \times p$ full rank matrix (its columns are linearly independent), $\mathbf{y}$ and $\mathbf{e}$ are $n \times 1$, and β is $p \times 1$. Assume $\mathbf{X}$ is a fixed (non-random) matrix, $E(\mathbf{e}) = \mathbf{0}$, and $cov(\mathbf{e}) = \sigma^2\mathbf{I}$. The least squares projection matrix $\mathbf{H}$ is an $n \times n$ matrix, the "hat" matrix, of the form $\mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T}$.

For each of the following statements, verify that it is true, or state why it is false:

a) $\hat{y} = \mathbf{X}\hat{\beta} = \mathbf{H}\mathbf{y}$

True. Recall that $\hat{\beta} = (\mathbf{X^T X})^{\mathbf{T}}\mathbf{X^T y}$. Substituting into the expression we have:

$$\mathbf{Hy} = \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T y} = \mathbf{X}\hat{\beta}$$

b) $\hat{\beta} = \mathbf{H}\beta$

False: Where do I start? First of all the dimensions are not right because $\mathbf{H}$ is $n \times n$ while β is $p \times 1$, so they cannot be multiplied this way. Second, $\hat{\beta}$ is an observable statistic that depends on the data only, while β is an unobservable parameter.

c) $\mathbf{HH} = \mathbf{H}$

True:

$$\mathbf{HH} = \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T X}(\mathbf{X^T X})^{-1}\mathbf{X^T} = \mathbf{XI}(\mathbf{X^T X})^{-1}\mathbf{X^T} = \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T} = \mathbf{H}$$

d) $(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}$

True:

$$(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{IX} - \mathbf{HX} = \mathbf{X} - \mathbf{X}(\mathbf{X^T X})^{-1}\mathbf{X^T X} = \mathbf{X} - \mathbf{X} = \mathbf{0}$$

e) $cov(\mathbf{y} - \hat{\mathbf{y}}) = \sigma^2(\mathbf{I} - \mathbf{H})$

True: First, $cov(\mathbf{y}) = cov(\mathbf{e}) = \sigma^2\mathbf{I}$. So, using Part a),

$$cov(\mathbf{y} - \hat{\mathbf{y}}) = cov((\mathbf{I} - \mathbf{H})\mathbf{y}) = (\mathbf{I} - \mathbf{H})(\sigma^2\mathbf{I})(\mathbf{I} - \mathbf{H})^T = \sigma^2(\mathbf{I} - \mathbf{H} - \mathbf{H^T} + \mathbf{HH^T})$$

Using the symmetry of $\mathbf{H}$ and Part c), this reduces to

$$cov(\mathbf{y} - \hat{\mathbf{y}}) = \sigma^2(\mathbf{I} - \mathbf{H} - \mathbf{H} + \mathbf{HH}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

5. A study was conducted to compare antibiotic (drug) treatment with placebo (no drug) for a certain disease. The variables are:

$$x_{i1} = \text{Pretreatment condition score}$$
$$x_{i2} = \begin{cases} 1, & \text{Treated with drug;} \\ 0, & \text{Treated with placebo (no drug).} \end{cases}$$
$$y_i = \text{Post-Treatment condition score (condition after treatment)}$$

The following model was fit using the `lm` function in **R**: $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$, $i = 1, 2, \ldots, n$, where the working assumptions are that the errors e_i are independently distributed as $N(0, \sigma^2)$. Some results are below.

```
##
## Call:
## lm(formula = PostTreatment ~ Pretreatment + Drug, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4110 -2.3897 -0.5214  1.6708  8.5890
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.4429     2.4216  -0.183   0.8562
## Pretreatment   0.9878     0.1611   6.132  1.5e-06 ***
## Drug          -3.3896     1.6100  -2.105   0.0447 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.931 on 27 degrees of freedom
## Multiple R-squared:  0.6762, Adjusted R-squared:  0.6522
## F-statistic: 28.19 on 2 and 27 DF,  p-value: 2.446e-07
```

a) What was the overall sample size, n?

Residual degrees of freedom $= 27 = n - 3$, so $n = 30$.

b) In the model summary, a t value is reported for `Drug`. State the null hypothesis and alternative hypothesis for this test, expressed in terms of the mathematical parameters (β_0, β_1, β_2, σ^2). Is the null hypothesis rejected at level 0.05?

$$H_0 : \beta_2 = 0; \qquad H_a : \beta_2 \neq 0$$

The p-value$= 0.0447 < 0.05$, so we reject H_0.

c) At the end of the model summary, an F test result is reported. State the null hypothesis and alternative hypothesis for this test, expressed in terms of the mathematical parameters (β_0, β_1, β_2, σ^2). Is the null hypothesis rejected at level 0.05?

$$H_0 : \beta_1 = \beta_2 = 0; \qquad H_a : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0$$

The p-value $= 2.5x10^{-7} << 0.05$, so, yes, H_0 is rejected.

d) Based on the fitted model, estimate the expected post-treatment condition score for a new patient with pre-treatment score of 10 if they receive the drug. Also compute the post-treatment condition score if they received the placebo.

$$\text{Given Drug: } \hat{\mu} = -0.443 + 0.988 * 10 - 3.39 = 6.05$$

$$\text{Given Placebo: } \hat{\mu} = -0.443 + 0.988 * 10 = 9.44$$

e) What does the $fit component in the output below tell us? Explain it briefly.

```
predict(mod1, newdata=data.frame(Pretreatment=12, Drug=0),
        se.fit=TRUE, interval="prediction", level=0.90)
```

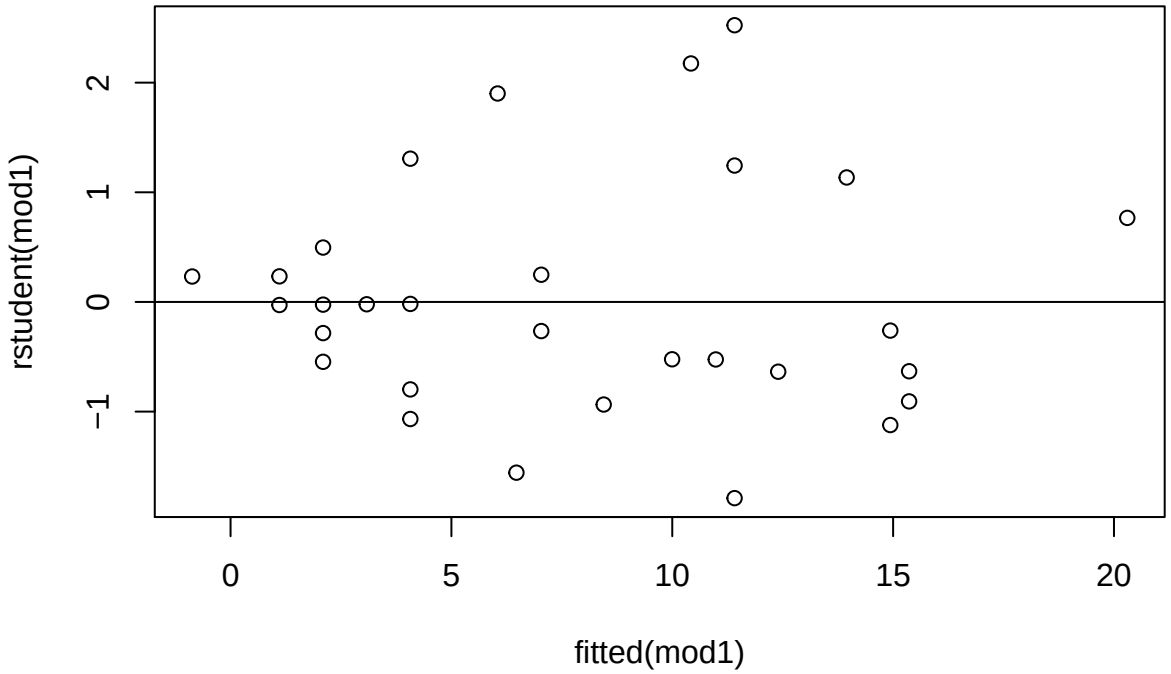
```
## $fit
##        fit      lwr      upr
## 1 11.41096 4.383872 18.43804
##
## $se.fit
## [1] 1.251574
##
## $df
## [1] 27
##
## $residual.scale
## [1] 3.931175
```

$fit['fit'] is the predicted post treatment score of 11.4 for a patient with a pretreatment score of 12 who is given placebo.

The 'lwr' and 'upr' parts give a 90% prediction interval for such a patient. The interval ranges from 4.38 to 18.43.

6. This problem considers the same data and model as in Problem 5.

a) A scatter plot of studentized residuals versus fitted values is shown below.

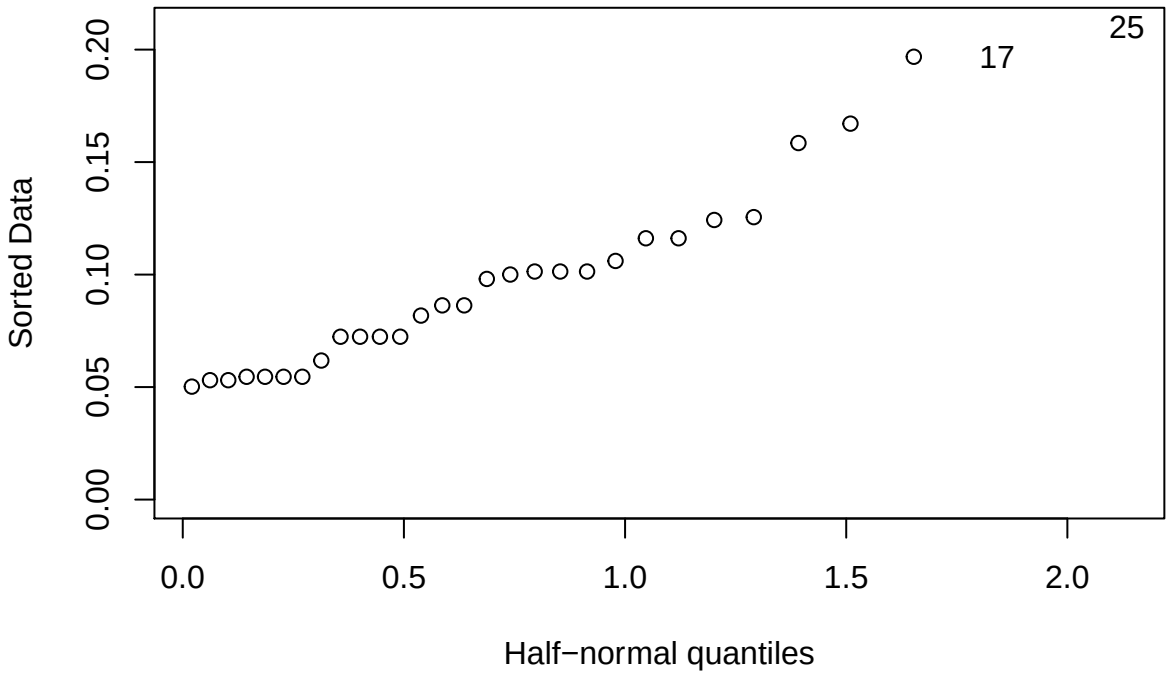


There is a hint of increasing vertical spread in the graph as we move from left to right. If that effect is real, what does it suggest about the model assumptions we have made? Describe briefly.

It suggests that the error variance is not constant, but instead depends on the mean.

b) What are the values being graphed below, and what do they tell us?

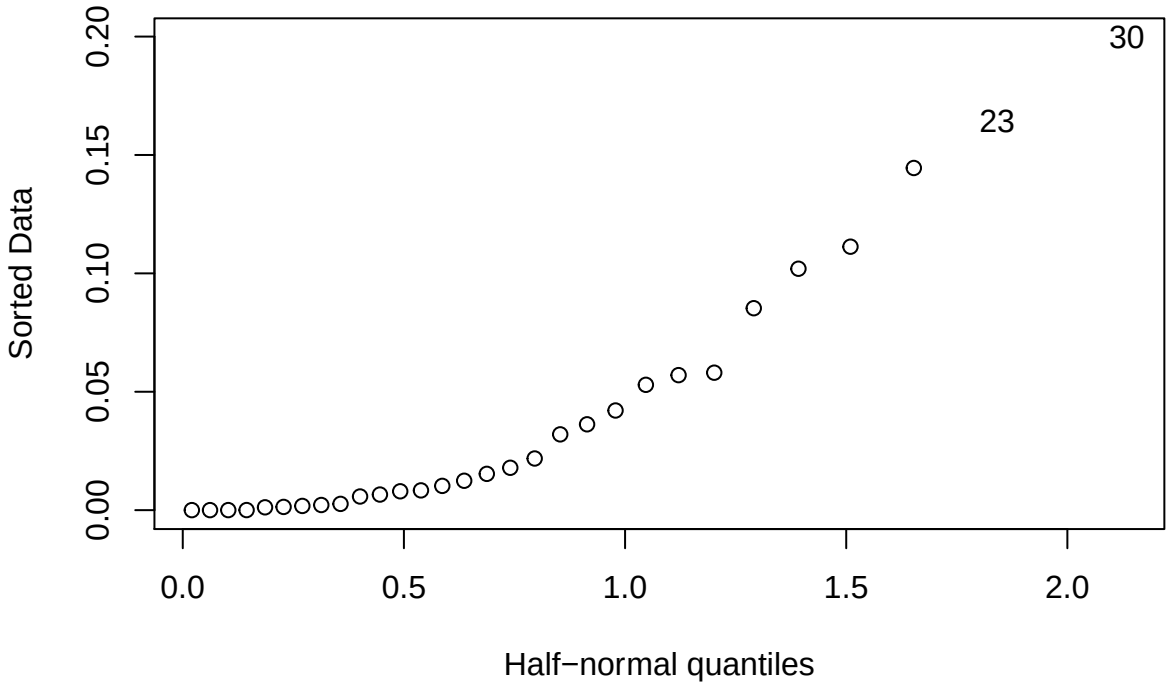
```
library(faraway)
halfnorm(influence(mod1)$hat)
```



These are the "leverages" or diagonals of the hat matrix for the different observations. They are sorted so we can tell which observations have the most *potential* to influence the model.

c) What are the values being graphed below, and what do they tell us?

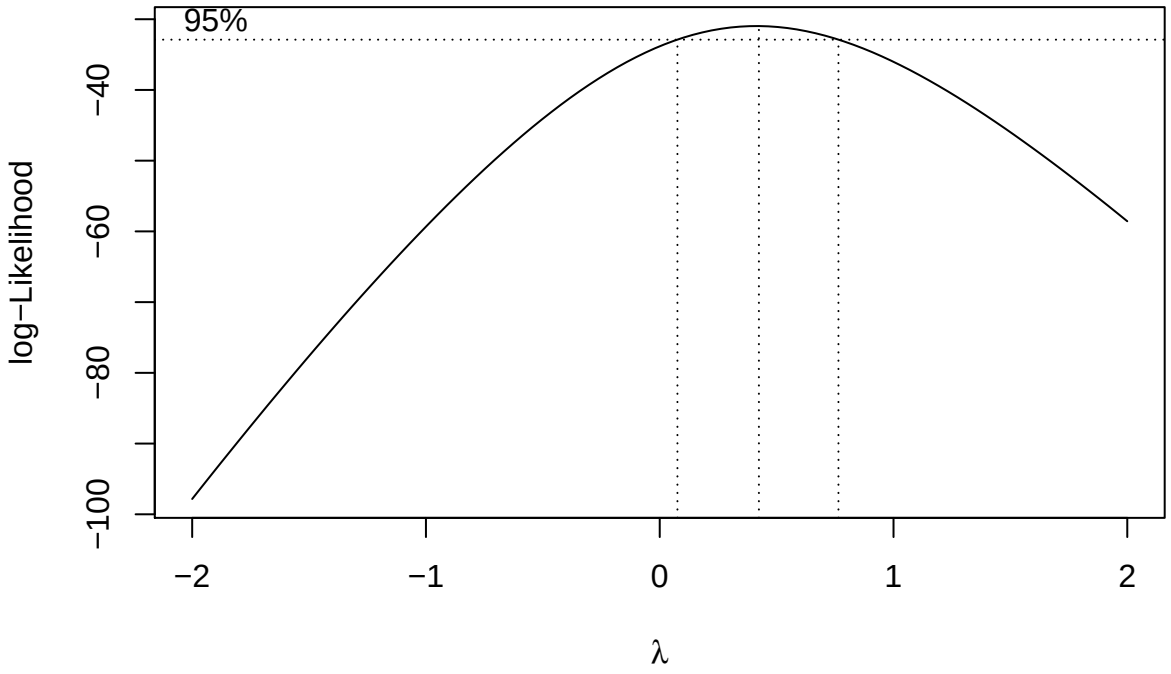
```
halfnorm(cooks.distance(mod1))
```



These are the Cook's Distance measures for the different observations. They tell which observations are most influential on the model in terms of causing the largest changes the fitted values and coefficients if they are removed. They are sorted so we can tell which observations have the most *influence* the model.

7. Continuing with the data and variables defined in Problem 5, consider the following **R** code and results:

```
library(MASS)
lambdas = boxcox(PostTreatment+1 ~ Pretreatment+Drug, data=df)
```




```
attributes(lambdas)
```

```
## $names
## [1] "x" "y"
```

```
lambdas$x[lambdas$y==max(lambdas$y)]
```

```
## [1] 0.4242424
```

a) Write out the class of regression models being fit here, expressed in terms of our original variables x_{1i}, x_{2i}, y_i, e_i, and the parameters β_0, β_1, β_2, σ^2, and λ.

$$\frac{(y_i+1)^\lambda - 1}{\lambda} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i$$

where the errors are independently distributed as $N(0, \sigma^2)$.

b) What is the estimated numerical value of λ? Also, give an approximate 95% confidence interval for λ.

The maximum likelihood estimate is $\hat{\lambda} = 0.424$. Based on the log-likelihood curve, the approximate range of the 95% confidence interval is (0.05, 0.80).

c) Do these results indicate that we should modify the linear model we fit in Problem 5? If so what do you recommend?

Yes, because the no-transformation value, $\lambda = 1$, is rejected at level 0.05 because it is outside of the 95% confidence interval. The results suggests fitting a linear model to a transformation of PostTreatment+1 rather than PostTreatment itself. A square root transformation is close to the MLE and could be a reasonable choice for improving the model fit. (The +1 is to avoid zeros in the scores).