STAT 425 Exam 2 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

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, 2;\ j = 1, 2, \ldots, 10.$$

Here y_{ij} denotes the 14-day weight gain for the jth chick on Diet i with $i = 1$ for Diet A and $i = 2$ for Diet B. The working model is that the errors are independently normally distributed with mean zero and variance σ^2.

a) Suppose the sample mean responses for the two diet groups are $\bar{y}_A = 101.2$ and $\bar{y}_B = 123.7$. Using the reference category constraint with Diet A as the reference category, calculate the least squares estimates of μ, α_1 and α_2.

$$\hat{\mu} = \bar{y}_A = 101.2$$
$$\hat{\alpha}_1 = 0$$
$$\hat{\alpha}_2 = \bar{y}_B - \bar{y}_A = 123.7 - 101.2 = 22.5$$

b) Calculate the between group sum of squares $FSS = \sum_{i=1}^{2} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$.

$$n_1 = n_2 = 10 \qquad \bar{y}_{1\cdot} = 101.2 \qquad \bar{y}_{2\cdot} = 123.7$$
$$\bar{y}_{\cdot\cdot} = \frac{10 * 101.2 + 10 * 123.7}{20} = \frac{101.2 + 123.7}{2} = 112.45$$
$$FSS = 10 * (101.2 - 112.45)^2 + 10 * (123.7 - 112.45)^2 = 2 * 10 * 11.25^2 = 2531.25$$

c) How many degrees of freedom does FSS have?

$$2 - 1 = 1$$

d) Suppose $\sum_{i=1}^{2} \sum_{j=1}^{10} (y_{ij} - \hat{y}_{ij})^2 = 49.0$. Calculate the value of the F-statistic for testing the null hypothesis $H_0 : \mu_1 = \mu_2 = 0$, where μ_1 is the mean response for Diet A, and μ_2 is the mean response for Diet B.

$$F = \frac{FSS/1}{RSS/(20-2)} = \frac{2531.25}{49/18} = 929.85$$

2. A study was conducted to compare three drug treatments for a certain disease, with drugs labeled A, B and C. For each subject `Pretreatment` is a condition score before treatment. The response `PostTreatment` is the condition score after the treatment regimen. The goal is to determine whether there is a difference between the drugs in improving post-treatment condition, adjusting for the effect of pre-treatment condition.

An analysis of covariance model was fit including the interactions between `Drug` and `Pretreatment`, and the sequential anova table is below.

```
## Analysis of Variance Table
##
## Response: PostTreatment
##                    Df Sum Sq Mean Sq F value    Pr(>F)
## Pretreatment        1 802.94  802.94 48.4726 3.366e-07 ***
## Drug                2  68.55   34.28  2.0692    0.1482
## Pretreatment:Drug   2  19.64    9.82  0.5930    0.5606
## Residuals          24 397.56   16.56
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a) What is the overall sample size, n?

$$5 + 24 + 1 = 30$$

b) What hypothesis is being tested by the following row in the table, and what do you conclude from the result?

```
Pretreatment:Drug 2 19.64 9.82 0.5930 0.5606
```

Equivalent statements:

H_0 : No interaction between Pretreatment and Drug effects

H_0 : The additive model is adequate: `PostTreatment` ~ `Pretreatment + Drug`

c) Based on the sequential anova results what is the best model for these data:

`PostTreatment ~ 1`

`PostTreatment ~ Pretreatment`

`PostTreatment ~ Pretreatment + Drug`

`PostTreatment ~ Pretreatment + Drug + Pretreatment:Drug`

Explain why.

`PostTreatment ~ Pretreatment`. Reason: stepping backward from the full interaction model, the interaction terms are not significant with the two main effects in the model, and, dropping the interaction, the main effect for `Drug` is not significant with `Pretreatment` in the model.

d) Write out the model formula (in R syntax) for the model that corresponds to three parallel regression lines for PostTreatment versus Pretreatment for the three Drug groups.

`PostTreatment ~ Pretreatment + Drug`

e) Consider the following notation to express the variables in the data mathematically for the ith subject:

y_i is the PostTreatment score,

x_i is the Pretreatment score,

$z_{i1} = 1$ if Drug A and $z_{i1} = 0$ if not Drug A,

$z_{i2} = 1$ if Drug B and $z_{i2} = 0$ if not Drug B,

$z_{i3} = 1$ if Drug C and $z_{i3} = 0$ if not Drug C,

and e_i is the error term. Using this notation, write out a valid, full rank mathematical form of the model corresponding to the R formula:

`PostTreatment ~ Pretreatment + Drug + Pretreatment:Drug`

Use expressions like β_0, β_1 etc. for the coefficients of the model.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_{i2} + \beta_3 z_{i3} + \beta_4 x_i z_{i2} + \beta_5 x_i z_{i3} + e_i, \quad i = 1, 2, \ldots, 30$$

Note that we only use two of the three indicator variables to distinguish the three Drug categories. We can tell it's Drug A if $z_{i2} = z_{i3} = 0$. So we make Drug A the reference value, and β_2 and β_3 are incremental effects of Drugs B and C, respectively, versus Drug A. The interactions are coded as products of the Pretreatment and Drug variables.

3. A cubic polynomial was fit using the crossx variable as the response and energy as the predictor. The sequential ANOVA table for this fitted model is as follows:

```
## Analysis of Variance Table
##
## Response: crossx
##              Df  Sum Sq Mean Sq   F value    Pr(>F)
## energy        1 272.216 272.216 1265.6028 3.289e-08 ***
## I(energy^2)   1  10.980  10.980   51.0492 0.0003789 ***
## I(energy^3)   1   0.622   0.622    2.8933 0.1398498
## Residuals     6   1.291   0.215
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

a) Based on the numerical results above, select one of the following options to describe the most appropriate model for this data set, and explain your choice:

1. A linear trend or simple linear regression of `crossx` on `energy`.
2. A quadratic polynomial model.
3. A cubic polynomial model.
4. We do not have enough information to select among the options above.

   2. A quadratic polynomial model. Reason: Starting from the bottom of the sequential anova table, the cubic term is not statistically significant, so we would drop it. The quadratic term `I(energy`^{2}`)` is highly significant so we keep it and stop.

b) In the sequential ANOVA table above, the F test corresponding to the quadratic term `I(energy`^{2}`)` is calculated as the ratio of two numbers

1. Numerator: A= (use a number with two digits after the decimal)
2. Denominator: B= (use a number with two digits after the decimal)

Give the numerator A, denominator B, and degrees of freedom for this F test.

$$A = 10.98 \text{ (I(energy}^2\text{) Mean Sq)}, \quad B = 0.215 \text{ (Residuals Mean Sq)}$$

$$df = 1 \text{ and } 6 \text{ (numerator and denominator)}$$

c) Consider the following notation to express the variables mathematically for the ith observation: y_i is the value for `crossx`, x_i is the value for `energy`, and e_i is the error. Using this notation, write out a valid mathematical form of the model corresponding to the cubic polynomial model. For the coefficients, use expressions like β_0, β_1, etc.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + e_i,\ i = 1, 2, \ldots, 10$$

4. Each of the following R function calls create a set of basis functions. For each, give the degrees of freedom and the total number of knots. Note - problem should have specified that the lm function will include the intercept if it's not already in the basis functions.

a) B-spline:

```
bs(year, df=6, intercept=TRUE)
```

df: 6

knots: 6-4=2

b) B-spline:

```
bs(year, df=8, intercept=FALSE)
```

df: 9 assuming the intercept will be included in the model, e.g. by `lm`

knots: 9-4 = 5

c) Natural Cubic Spline:

```
ns(year, df=8, intercept=TRUE)
```

df: 8

knots: 8-2=6

d) Natural Cubic Spline:

```
ns(year, df=10, intercept=FALSE)
```

df: 10+1=11

knots: 11-2=9

5. The following output summarizes the options for the first step in a backwards stepwise selection process applied to a linear model for data relating mean life expectancy for U.S. states to other demographics. The first column is the list of candidate variables for deletion. The RSS column shows the residual sum of squares for the model without the indicated variable, and the AIC column shows the AIC for that model.

```
## Single term deletions
##
## Model:
## Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
##     Frost + Area
##            Df Sum of Sq    RSS     AIC
## <none>                  23.297 -22.185
## Population  1    1.7472 25.044 -20.569
## Income      1    0.0044 23.302 -24.175
## Illiteracy  1    0.0047 23.302 -24.174
## Murder      1   23.1411 46.438  10.305
## HS.Grad     1    2.4413 25.738 -19.202
## Frost       1    1.8466 25.144 -20.371
## Area        1    0.0011 23.298 -24.182
```

a) Which, if any, variable would be removed if we use AIC as the selection criterion for backward stepwise regression? Explain why.

`Area`: dropping this variable gives the lowest AIC of -24.182

b) What is the AIC value for the model that includes all variables?

-22.185

c) The full model we started with has smaller RSS than any of the candidate models obtained by dropping one variable. Does this mean that the full model is actually the best option? Explain why or why not.

No. RSS only measures fit, not complexity. We can driving down the RSS by adding more variables, but this leads to over-fitting the particular data set. The resulting model would not be a good predictive model for new data. AIC is more reliable because it prevents over-fitting by penalizing complexity.

6. Several questions about regularized regression.

a) When we use principal components regression, it is always best to use all principal components of the design matrix $\mathbf{X}$ for regression. True or False? Explain.

> False. This would be equivalent to using all of the variables in $\mathbf{X}$. We will obtain better predictions by reducing to a smaller set of principal components that account for a large percentage of the variation in $\mathbf{X}$.

b) Consider the following output after computing the principal components of predictor variables:

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6    PC7
## Standard deviation     1.7548 1.2739 1.0025 0.74634 0.58222 0.50886 0.3713
## Proportion of Variance 0.4399 0.2318 0.1436 0.07958 0.04843 0.03699 0.0197
## Cumulative Proportion  0.4399 0.6717 0.8153 0.89488 0.94331 0.98030 1.0000
```

According to this output, how many principal components are needed to account for at least 90% of the variation in the predictor variables? Give the actual percentage of variation accounted for by this choice.

> 5 principal components. From the "Cumulative Proportion" row the first 5 principal components account for **94.3%** of the variation, whereas the first 4 account for slightly less than 90%.

c) Lasso regression minimizes the residual sum of squares of a regression model subject to the constraint that $\sum_{j=1}^{p} |\beta_j| \leq t$. Which of the following methods could be used to select the value for t:

1. Maximum likelihood
2. Least squares
3. Weighted least squares
4. Cross-validation
5. None of the above

> 4. Cross-validation. This provides an unbiased way to estimate which constrained model gives the best out-of-sample predictions.

7. The following results are from fitting a one-way anova model to data relating blood coagulation times to diet.

```
##
## Call:
## lm(formula = coag ~ diet, data = coagulation)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5.00  -1.25   0.00   1.25   5.00
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.100e+01  1.183e+00  51.554  < 2e-16 ***
## dietB       5.000e+00  1.528e+00   3.273 0.003803 **
## dietC       7.000e+00  1.528e+00   4.583 0.000181 ***
## dietD       2.991e-15  1.449e+00   0.000 1.000000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.366 on 20 degrees of freedom
## Multiple R-squared:  0.6706, Adjusted R-squared:  0.6212
## F-statistic: 13.57 on 3 and 20 DF,  p-value: 4.658e-05
```

a) Based on these results, compute the estimated difference between the mean coagulation time for Diet B and the mean for Diet A. Also give the standard error for this difference if possible.

> We can see from the results that Drug A is the reference group, so this mean difference is estimated by the `dietB` coefficient. Estimate = 5.00, Standard Error = 1.53

b) Below is the analysis of variance table for the model above.

```
## Analysis of Variance Table
##
## Response: coag
##           Df Sum Sq Mean Sq F value    Pr(>F)
## diet       3    228    76.0  13.571 4.658e-05
## Residuals 20    112     5.6
```

State the null hypotheses that the F value is testing. Does the test reject the null hypothesis at level 0.05?

$$H_0 : \mu_A = \mu_B = \mu_C = \mu_D \text{ (all treatment group means are equal)}$$

p < .0001 so the test rejects the null hypothesis at level 0.05.

c) Based on the results below, determine which pairs of diet group means are significantly different from each other, controlling the family wise error rate at $\alpha = 0.05$.

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = g)
##
## $diet
##     diff         lwr       upr     p adj
## B-A    5   0.7245544  9.275446 0.0183283
## C-A    7   2.7245544 11.275446 0.0009577
## D-A    0  -4.0560438  4.056044 1.0000000
## C-B    2  -1.8240748  5.824075 0.4766005
## D-B   -5  -8.5770944 -1.422906 0.0044114
## D-C   -7 -10.5770944 -3.422906 0.0001268
```

The following pairs of mean differences have adjusted p-values < 0.05 and are therefore statistically significant:

$$B - A\,(> 0), \quad C - A\,(> 0), \quad D - B\,(< 0), \quad D - C\,(< 0)$$

Another way to think about this result is that the means are grouped as follows:

$$\{A, D\} < \{B, C\}$$

with significant differences between the two bracketed groups, and no significant differences within the bracketed groups.

8. Consider the following linear model assumptions.

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e}, \quad E(\mathbf{e}) = \mathbf{0}, \quad Cov(\mathbf{e}) = \sigma^2\mathbf{V},$$

where $\mathbf{V}$ is a diagonal matrix of the form

$$\mathbf{V} = \begin{pmatrix} v_1 & 0 & 0 & \cdots & 0 \\ 0 & v_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & v_n \end{pmatrix}$$

and the diagonal elements are all positive.

a) Let $\mathbf{W}$ be the diagonal matrix

$$\mathbf{W} = \begin{pmatrix} \frac{1}{\sqrt{v_1}} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{v_2}} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & 0 \\ 0 & 0 & 0 & \cdots & \frac{1}{\sqrt{v_n}} \end{pmatrix}$$

Show that the transformed response vector $\mathbf{Z} = \mathbf{WY}$ follows a modified linear model in which the error vector has mean $\mathbf{0}$ and covariance $\sigma^2\mathbf{I}_n$.

Using the model equation we have

$$\mathbf{Z} = \mathbf{WY} = \mathbf{WX}\beta + \mathbf{We} = \mathbf{X}^*\beta + \mathbf{e}^*,$$

where $\mathbf{X}^* = \mathbf{WX}$ and $\mathbf{e}^* = \mathbf{We}$. Furthermore, $E(\mathbf{e}^*) = WE(\mathbf{e}) = 0$ and $Cov(\mathbf{e}^*) = \sigma^2\mathbf{WVW^T} = \sigma^2\mathbf{I}$.

b) The weighted sum of squared residuals for the original model has the form

$$RSS_w(\beta) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)^2/v_i = (\mathbf{Y} - \mathbf{X^T}\beta)\mathbf{WW}(\mathbf{Y} - \mathbf{X^T}\beta)$$

Show that minimizing $RSS_w(\beta)$ as a function of β is the same as miminizing the ordinary (unweighted) residual sum of squares as a function of β for the modified linear model for $\mathbf{Z}$.

First note that $\mathbf{W^T} = \mathbf{W}$ and $\mathbf{WW} = \mathbf{V}^{-1}$. The OLS residual sum of squares for the transformed model in a) is given by

$$\begin{aligned} RSS^*(\beta) &= (\mathbf{Z} - \mathbf{X}^*\beta)^T(\mathbf{Z} - \mathbf{X}^*\beta) \\ &= (\mathbf{WY} - \mathbf{WX}\beta)^T(\mathbf{WY} - \mathbf{WX}\beta) \\ &= (\mathbf{W}(\mathbf{Y} - \mathbf{X}\beta))^T(\mathbf{W}(\mathbf{Y} - \mathbf{X}\beta)) \\ &= (\mathbf{Y} - \mathbf{X}\beta)^T\mathbf{W}^T\mathbf{W}(\mathbf{Y} - \mathbf{X}\beta) \\ &= (\mathbf{Y} - \mathbf{X}\beta)^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\beta) = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\beta)^2/v_i \end{aligned}$$

Therefore, minimizing $RSS_w(\beta)$ is the same as minimizing $RSS^*(\beta)$.

STAT 425 Assignment 1

Due Monday, February 8, 11:59pm. Submit through Moodle.

Name: (insert your name here)

Netid: (insert)

Submit your computational work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown.

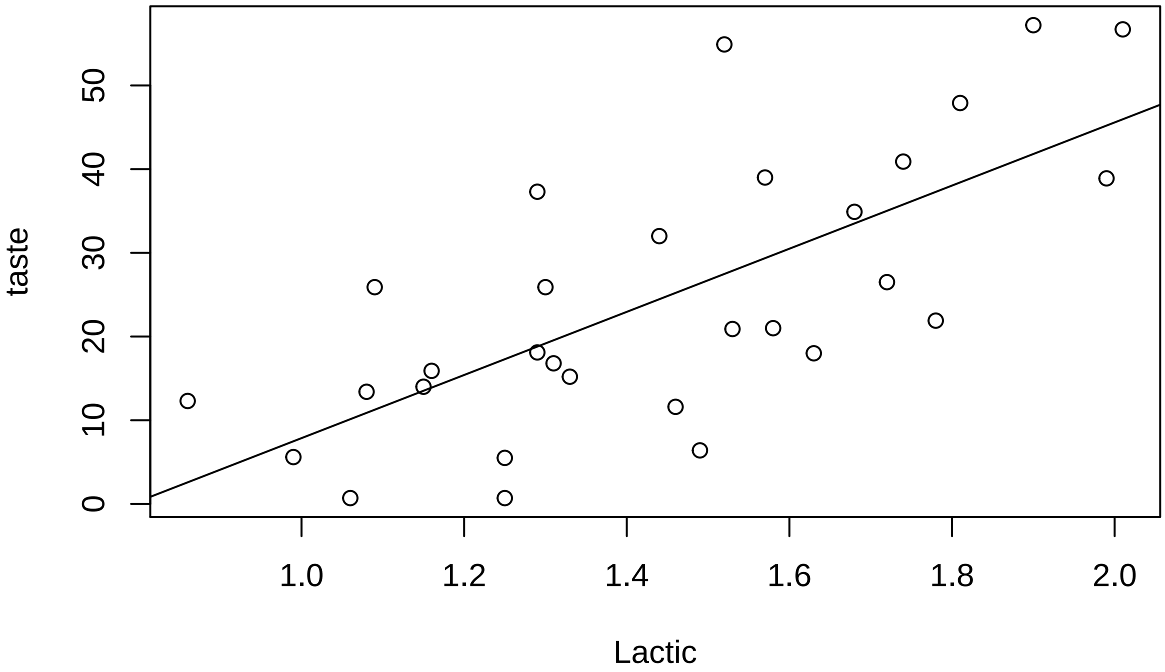
Problem 1

Thirty samples of cheddar cheese were analyzed for their content of acetic acid, hydrogen sulfide and lactic acid. A panel of judges tasted each sample and scored them, and the average taste score for each sample was recorded. The data are available as the data frame 'cheddar' in the **faraway** library. After loading the library enter 'help(cheddar)' for more information.

a) Make a scatter plot of 'taste' versus 'Lactic' and include the least squares regression line on the graph. Comment on whether the graph appears consistent with data that follow a linear model.

Answer:

```r
library(faraway)
attach(cheddar)
plot(Lactic, taste) +
  abline(lm(taste ~ Lactic))
```



```
## integer(0)
```

```
#detach(cheddar)
```

The data points follow a fairly strong linear model.

b) Obtain and display the summary of the least square fitted model, including coefficient estimates, standard errors, t-values and p-values. Is there is a statistically significant association between lactic acid content and the average taste score, using a significance level of $\alpha = 0.05$? Explain based on your results, making clear what information from the results you are using.

```
fit = lm(taste ~ Lactic)

#summary
summary = summary(fit)

#coefficient estimates, standard errors, t-values, and p-values.
summary$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -29.85883  10.582319 -2.821577 8.690703e-03
## Lactic       37.71995   7.186396  5.248799 1.405117e-05
```

With a p-value of 1.405e-05, we have enough evidence to conclude at the .05 significance level that there is a statistically significant association between lactic acid content and average taste score. We can reject H0 and conclude the linear coefficient is not 0.

Answer:

c) In R, the 'cor' function can compute the sample correlation coefficient between two variables in a data set. Compute the **squared** correlation between 'taste' and 'Lactic'. Verify that this is numerically equal to R^2 for the model. (Note: to refer to a variable within a data frame use the dataframe$variable syntax.)

```
cor(taste, Lactic)^2
```

```
## [1] 0.4959486
```

```
summary$r.squared
```

```
## [1] 0.4959486
```

```
cor(taste, Lactic)^2 - summary$r.squared
```

```
## [1] 5.551115e-17
```

Answer:

d) Compute a 95% confidence interval for the coefficient of 'Lactic' in the model.

```
confint(fit, 'Lactic', level = .95)
```

```
##           2.5 %   97.5 %
## Lactic 22.99928 52.44061
```

Answer:

e) Compute a 95% confidence interval for the mean taste value expected for a cheddar cheese sample with lactic acid concentration of 2.0.

Answer:

```
fit2 = lm(taste ~ Lactic)
predict(fit, data.frame(Lactic = 2.0), interval = "confidence",level = .95)
```

```
##        fit      lwr      upr
## 1 45.58106 36.26624 54.89588
```

Problem 2

The simple regression through the origin model has the form

$$y_i = \beta_1 x_i + e_i, \quad i = 1, 2, \ldots, n,$$

where the standard assumptions are that $E(e_i) = 0$, $\text{var}(e_i) = \sigma^2$, and $\text{cov}(e_i, e_j) = 0$ if $i \neq j$. The least squares estimate of β_1 minimizes the residual sum of squares,

$$RSS(\beta_1) = \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2$$

as a function of β_1.

a) Take the derivative of $RSS(\beta_1)$ with respect to β_1, set the derivative to zero. Solve the resulting equation algebraically to obtain the formula for the estimate, $\hat{\beta}_1$.

Answer: Taking the partial derivative of $RSS(\beta_1)$ with respect to β_1 yields

$$\frac{\partial}{\partial \beta_1} RSS(\beta_1) = \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (y_i - \beta_1 x_i)^2 = 2 \sum_{i=1}^{n} x_i (y_i - \beta_1 x_i).$$

Setting the derivative to zero yields the equation

$$\sum_{i=1}^{n} x_i y_i = \beta_1 \sum_{i=1}^{n} x_i^2,$$

which implies that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

b) Use your formula form Part a) to show that $\hat{\beta}_1$ is an unbiased estimator of β_1 under the standard assumptions.

Answer: Note that only y_i is random and x_i is fixed for $i = 1, \ldots, n$. Taking expectation of $\hat{\beta}_1$ results in

$$\begin{aligned} E(\hat{\beta}_1) = \frac{\sum_{i=1}^{n} x_i E(y_i)}{\sum_{i=1}^{n} x_i} &= \frac{\sum_{i=1}^{n} x_i E(\beta_1 x_i + e_i)}{\sum_{i=1}^{n} x_i^2} \\ &= \frac{\beta_1 \cdot \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} && (E(e_i) = 0) \\ &= \beta_1, \end{aligned}$$

which shows that $\hat{\beta}_1$ is an unbiased estimator of β_1.

c) Show that

$$\operatorname{var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$

Answer: Through using ordinary variance operator, we can obtain

$$\operatorname{var}(\hat{\beta}_1) = \operatorname{var}\left(\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}\right) = \frac{\sum_{i=1}^{n} x_i^2 \cdot \operatorname{var}(y_i)}{\left(\sum_{i=1}^{n} x_i^2\right)^2},$$

where we used the fact that $\operatorname{cov}(e_i, e_j) = 0$ for $i \neq j$ and $\operatorname{cov}(y_i, y_j) = \operatorname{cov}(e_i, e_j)$ since x_i is fixed. Because $\operatorname{var}(y_i) = \operatorname{var}(\beta_1 x_i + e_i) = \operatorname{var}(e_i) = \sigma^2$, we can obtain

$$\operatorname{var}(\hat{\beta}_1) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{\left(\sum_{i=1}^{n} x_i^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2},$$

which finishes the proof.

d) Under the standard assumptions find $E(y_1)$ and $\text{var}(y_1)$.

Answer: Since $E(e_1) = 0$ and $E(e_1) = \sigma^2$,

$$E(y_1) = E(\beta_1 x_1 + e_1) = \beta_1 x_1 + E(e_1) = \beta_1 x_1$$
$$\text{var}(y_1) = \text{var}(\beta_1 x_1 + e_1) = \text{var}(e_1) = \sigma^2.$$

e) Under the standard assumptions find expressions for $E(\hat{y}_1)$ and $\text{var}(\hat{y}_1)$.

Answer: Since $\hat{y}_1 = \hat{\beta}_1 x_1$,

$$E(\hat{y}_1) = E(\hat{\beta}_1 x_1) = x_1 E(\hat{\beta}_1) = \beta_1 x_1,$$

where we used the fact that $\hat{\beta}_1$ is unbiased from b). Next,

$$\text{var}(\hat{y}_1) = \text{var}(\hat{\beta}_1 x_1) = x_1^2 \cdot \text{var}(\hat{\beta}_1) = \frac{\sigma^2 x_1^2}{\sum_{i=1}^n x_i^2},$$

where we use the expression for $\text{var}(\hat{\beta}_1)$ from c).

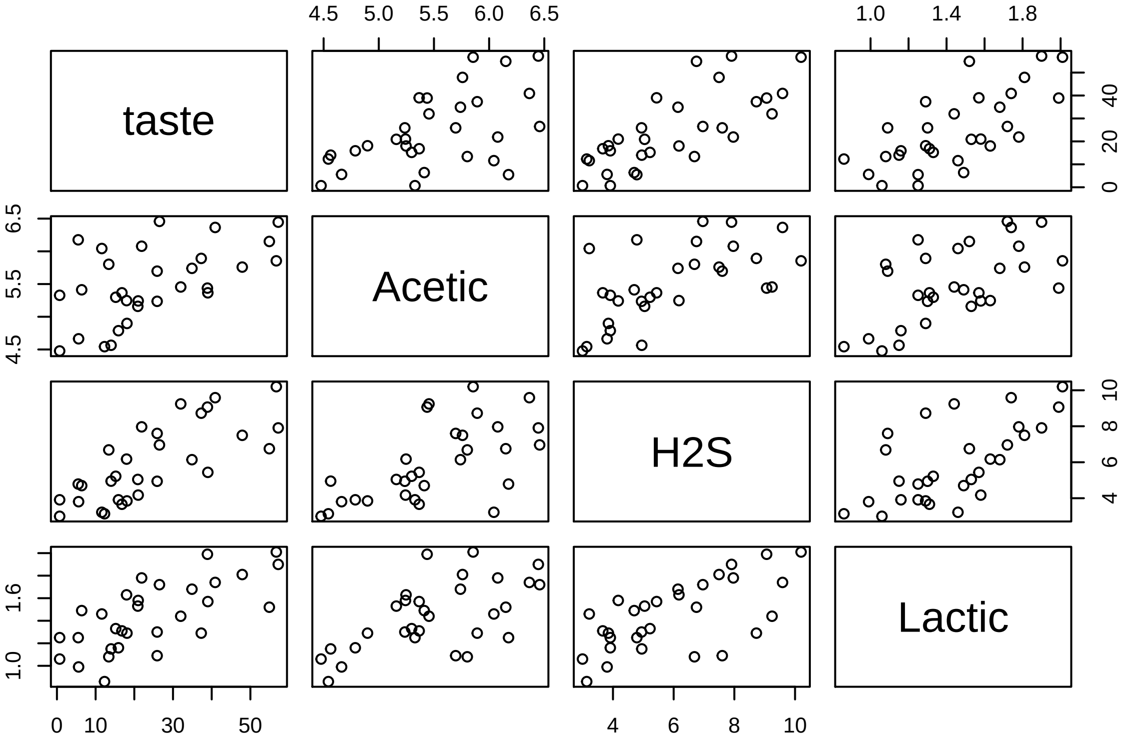
Problem 3:

This problem refers again to the ‘cheddar’ data described in Problem 1.

a) Make a ‘pairs’ plot of the data, i.e., a matrix of all the pairwise scatter plots between variables.

Answer:

```r
library("faraway")
plot(cheddar)
```



b) Fit a multiple linear regression model with 'taste' as the response and the three chemical constituent concentrations as the predictors. Display a summary of your fitted model. Note: the 'lm' function can fit a multiple linear regression model using a formula of the form 'y ∼ x1 + x2 + . . . + xp'.

Answer:

```
model = lm(taste~Acetic + H2S + Lactic, data = cheddar)
summary(model)
```

```
##
## Call:
## lm(formula = taste ~ Acetic + H2S + Lactic, data = cheddar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## Acetic        0.3277     4.4598   0.073  0.94198
## H2S           3.9118     1.2484   3.133  0.00425 **
## Lactic       19.6705     8.6291   2.280  0.03108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 10.13 on 26 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6116
## F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

c) Report the values of the regression coefficients for associated with the predictors.

Answer:

```
coef(model)
```

```
## (Intercept)      Acetic         H2S      Lactic
## -28.8767696   0.3277413   3.9118411  19.6705434
```

d) Which of the predictor variables have statistically significant coefficients, rejecting the null hypothesis that the coefficient is zero, at the 5% level of significance? Explain.

Answer: H2S and Lactic. By checking the summary of fitted model, we can see these two are the variable with p value smaller then 5%.

e) Compute an estimate of the average taste score for a cheddar sample with Acetic= 5.5, H2S=5.0, Lactic=1.5.

Answer:

```
predict(model, newdata= data.frame(Acetic= 5.5, H2S=5.0, Lactic=1.5))
```

```
##        1
## 21.99083
```

STAT 425 Assignment 2

Due Monday, February 22, 11:59pm. Submit through Moodle.

Name: (insert your name here)

Netid: (insert)

Submit your computational work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown.

Problem 1

In this problem we have data of the form $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. We construct a new variable z_i from the x_i values by subtracting their sample mean, so $z_i = x_i - \bar{x}$ for $i = 1, 2, \ldots, n$, where $\bar{x}_i = n^{-1}\sum_{i=1}^n x_i$. We consider two models:

$$\begin{aligned} &\text{Model 1:} \qquad y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, \ldots, n \\ &\text{Model 2:} \qquad y_i = \alpha_0 + \alpha_1 z_i + e_i, \quad i = 1, 2, \ldots, n \end{aligned}$$

Here the x_i values are considered a fixed set of numbers, and the errors e_i are considered to be uncorrelated random variables with $E(e_i) = 0$ and $\text{var}(e_i) = \sigma^2$.

a) By subtracting one model from the other and averaging, show that $\alpha_0 = \beta_0 + \beta_1\bar{x}$.

Answer: Since $z_i = x_i - \bar{x}$, by subtracting one model from the other yields

$$\beta_0 + \beta_1 x_i = \alpha_0 + \alpha_1(x_i - \bar{x}), \quad i = 1, 2, \ldots, n.$$

Summing the above equation for $i = 1, \ldots, n$ and taking average yields

$$\beta_0 + \beta_1\bar{x} = \alpha_0,$$

which follows from the fact that

$$\frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}) = \bar{x} - \bar{x} = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^n x_i = \bar{x}.$$

It finishes the proof.

(here or indicate where it is in the attached pdf file.)

b) Substituting $\alpha_0 = \beta_0 + \beta_1\bar{x}$ into Model 2, show that $(\alpha_1 - \beta_1)(x_i - \bar{x}) = 0$, for $i = 1, 2, \ldots, n$. Assume $\sum_{i=1}^n (x_i - \bar{x})^2 > 0$, i.e., the x_i values are not all the same. Show why this implies $\alpha_1 = \beta_1$.

Answer: Substituting $\alpha_0 = \beta_0 + \beta_1\bar{x}$ into Model 2 yields

$$\begin{aligned} y_i &= \beta_0 + \beta_1\bar{x} + \alpha_1 z_i + e_i \\ &= \beta_0 + \beta_1\bar{x} + \alpha_1(x_i - \bar{x}) + e_i, \quad i = 1, \ldots, n. \qquad (z_i = x_i - \bar{x}) \end{aligned}$$

Subtituting y_i in the right hand side of the above equation with $y_i = \beta_0 + \beta_1 x_i + e_i$ as in Model 1 shows that

$$\beta_0 + \beta_1 x_i + e_i = \beta_0 + \beta_1\bar{x} + \alpha_1(x_i - \bar{x}) + e_i, \quad i = 1, \ldots, n. \qquad (z_i = x_i - \bar{x})$$

Cancelling terms results in

$$\beta_1 x_i = \beta_1\bar{x} + \alpha_1(x_i - \bar{x}) \Rightarrow (\alpha_1 - \beta_1)(x_i - \bar{x}) = 0, \quad i = 1, \ldots, n.$$

It further implies

$$(\alpha_1 - \beta_1)^2(x_i - \bar{x})^2 = 0, \quad i = 1, \ldots, n.$$

Summing the above equation over $i = 1, \ldots, n$ yields

$$(\alpha_1 - \beta_1)^2 \sum_{i=1}^n (x_i - \bar{x})^2 = 0.$$

Because by assumption $\displaystyle\sum_{i=1}^n (x_i - \bar{x})^2 > 0$, the above equation implies $\alpha_1 = \beta_1$.

Now we use matrix notation and rewrite Model 2 as $\mathbf{y} = \mathbf{Z}\alpha + \mathbf{e}$ with:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2, \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{Z} = \begin{pmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_n \end{pmatrix} \qquad \alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \qquad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2, \\ \vdots \\ e_n \end{pmatrix}$$

c) Show

$$\mathbf{Z}^T\mathbf{Z} = \begin{pmatrix} n & 0 \\ 0 & S_{xx} \end{pmatrix}$$

where $S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$.

Answer: We can write

$$\mathbf{Z}^T\mathbf{Z} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_n \end{pmatrix} \begin{pmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_n \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n z_i \\ \sum_{i=1}^n z_i & \sum_{i=1}^n z_i^2 \end{pmatrix}.$$

Because

$$\begin{aligned}\sum_{i=1}^{n} z_i &= \sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0\\ \sum_{i=1}^{n} z_i^2 &= \sum_{i=1}^{n}(x_i - \bar{x})^2 = S_{xx},\end{aligned}$$

we can obtain

$$\mathbf{Z}^T\mathbf{Z} = \begin{pmatrix} n & 0 \\ 0 & S_{xx} \end{pmatrix}.$$

d) The LS estimate of the coefficient vector α solves the matrix equation:

$$\mathbf{Z}^T\mathbf{Z}\,\hat{\alpha} = \mathbf{Z}^T\mathbf{y}$$

Solve the equation algebraically to get simplified expressions for $\hat{\alpha}_1$ and $\hat{\alpha}_2$ expressed in terms of the original x_i, y_i and n values.

Answer: From c), for the equation $\mathbf{Z}^T\mathbf{Z}\hat{\boldsymbol{\alpha}} = \mathbf{Z}^T\mathbf{y}$, we can expand it to

$$\begin{aligned}\begin{pmatrix} n & 0 \\ 0 & S_{xx} \end{pmatrix}\begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix} &= \begin{pmatrix} 1 & 1 & \cdots & 1 \\ z_1 & z_2 & \cdots & z_n \end{pmatrix}\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}\\ &= \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n}(x_i - \bar{x})y_i \end{pmatrix}. \qquad (z_i = x_i - \bar{x})\end{aligned}$$

The above equation further implies

$$\begin{aligned}\hat{\alpha}_1 &= \frac{1}{n}\sum_{i=1}^{n} y_i,\\ \hat{\alpha}_2 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.\end{aligned}$$

The above two identities are the simplified expressions for $\hat{\alpha}_1$ and $\hat{\alpha}_2$.

e) Derive a simplified expressions for the entries of $\mathrm{cov}(\hat{\alpha})$, the 2×2 covariance matrix of $\hat{\alpha} = \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix}$. (Note that this still depends on the unknown σ^2.)

Answer: It suffices to compute $\text{var}(\hat{\alpha}_1)$, $\text{var}(\hat{\alpha}_2)$, and $\text{cov}(\hat{\alpha}_1, \hat{\alpha}_2)$. Note that by the assumption that e_i and e_j are uncorrelated for $i \neq j$,

$$\begin{aligned}
\text{var}(\hat{\alpha}_1) &= \text{var}(\bar{y}) = \frac{\sigma^2}{n}, \\
\text{var}(\hat{\alpha}_2) &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \text{var}(y_i)}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} = \sigma^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.
\end{aligned}$$

By linearity property of covariance,

$$\begin{aligned}
\text{cov}(\hat{\alpha}_1, \hat{\alpha}_2) &= \text{cov}\left(\frac{1}{n}\sum_{i=1}^{n} y_i, \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) \\
&= \frac{1}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \cdot \text{cov}\left(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n}(x_i - \bar{x})y_i\right).
\end{aligned}$$

Now, we further simplifies the last covariance term as below.

$$\begin{aligned}
\text{cov}\left(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n}(x_i - \bar{x})y_i\right) &= \sum_{i=1}^{n}(x_i - \bar{x})\text{var}(y_i) + \sum_{i \neq j}(x_j - \bar{x}) \cdot \text{cov}(y_i, y_j) \\
&= \sigma^2 \sum_{i=1}^{n}(x_i - \bar{x}) + \sum_{i \neq j}(x_j - \bar{x}) \cdot 0 \quad (\text{cov}(y_i, y_j) = \text{cov}(e_i, e_j) = 0) \\
&= 0. \qquad\qquad (\sum_{i=1}^{n}(x_i - \bar{x}) = 0)
\end{aligned}$$

Thus, it shows that $\text{cov}(\hat{\alpha}_1, \hat{\alpha}_2) = 0$. As a result,

$$\text{cov}(\hat{\boldsymbol{\alpha}}) = \begin{pmatrix} \sigma^2/n & 0 \\ 0 & \sigma^2 \Big/ \sum_{i=1}^{n}(x_i - \bar{x})^2 \end{pmatrix}.$$

Alternative solution:

From general results about LS regression,

$$\text{cov}(\hat{\boldsymbol{\alpha}}) = \sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1}.$$

Using Part c we obtain

$$\sigma^2(\mathbf{Z}^T\mathbf{Z})^{-1} = \sigma^2 \begin{pmatrix} n^{-1} & 0 \\ 0 & S_{xx}^{-1} \end{pmatrix}$$

Problem 2

This problem considers the `prostate` data in the library `faraway`. See `help(prostate)` for more information about the data set.

a) Fit a linear model with `lpsa` as the response and all the other variables as predictor variables. Display the model summary and identify all variables whose coefficient t values are statistically significant at the $\alpha = 0.05$ level. In each case, state what the null hypothesis is for the t test.

Answer:

```r
library(faraway)
fit_2a = lm(lpsa ~ ., data = prostate)
summary(fit_2a)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

The lcavol, lweight, and svi predictors are statistically significant at the $\alpha = 0.05$ significance level. The null hypothesis in each case is that the coefficient of a single predictor is 0 and that there is no linear association between this predictor and the lpsa response variable after adjusting for all other predictors.

b) Fit a reduced model with `lpsa` as the response but removing `lcp`, `gleason`, and `pgg45` as predictors in the model. Display the model summary, and provide both 90% and 95% confidence intervals for the coefficient of `lbph`. Do either or both of them include 0?

Answer:

```
fit_2b = lm(lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)
summary(fit_2b)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83505 -0.39396  0.00414  0.46336  1.57888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.95100    0.83175   1.143 0.255882
## lcavol       0.56561    0.07459   7.583 2.77e-11 ***
## lweight      0.42369    0.16687   2.539 0.012814 *
## age         -0.01489    0.01075  -1.385 0.169528
## lbph         0.11184    0.05805   1.927 0.057160 .
## svi          0.72095    0.20902   3.449 0.000854 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7073 on 91 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6245
## F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

```
confint(fit_2b, 'lbph', level = .95)
```

```
##             2.5 %    97.5 %
## lbph -0.003474551 0.2271544
```

```
confint(fit_2b, 'lbph', level = .90)
```

```
##             5 %      95 %
## lbph 0.01536969 0.2083102
```

Only the 95% confidence interval includes 0. This makes sense because the p-value of the t-test for the lbph predictor outputted in the summary is low enough to reject at the 90% confidence level, but not at the 95% confidence level.

c) Perform an F test comparing the model from Part b) as the null model, and the model from Part a) as the alternative model. Based on your calculations, does the test reject or accept the null hypothesis at the level $\alpha = 0.05$?

Answer:

```
anova(fit_2b, fit_2a)
```

```
## Analysis of Variance Table
##
## Model 1: lpsa ~ lcavol + lweight + age + lbph + svi
## Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     91 45.526
## 2     88 44.163  3    1.3625 0.905 0.4421
```

With a p-value of 0.4421, this test fails to reject the null hypothesis at the $\alpha = 0.05$ level. We accept the reduced model over the full model.

d) Here are the data for observation #32:

```
> prostate[32,]
      lcavol lweight age     lbph svi      lcp gleason pgg45    lpsa
32 0.1823216  6.1076  65 1.704748   0 -1.38629       6     0 2.00821
```

Pretending you did not know the value for `lpsa` for this observation, use your model from Part b) to compute a 95% prediction interval for observation 32 based on the predictors in the model. Does the interval include the value observed in the data?

Answer:

```
predict(fit_2b, newdata = prostate[32,], interval = "prediction", level = .95)
```

```
##         fit      lwr      upr
## 32 2.864524 1.244189 4.484858
```

```
prostate[32,]$lpsa
```

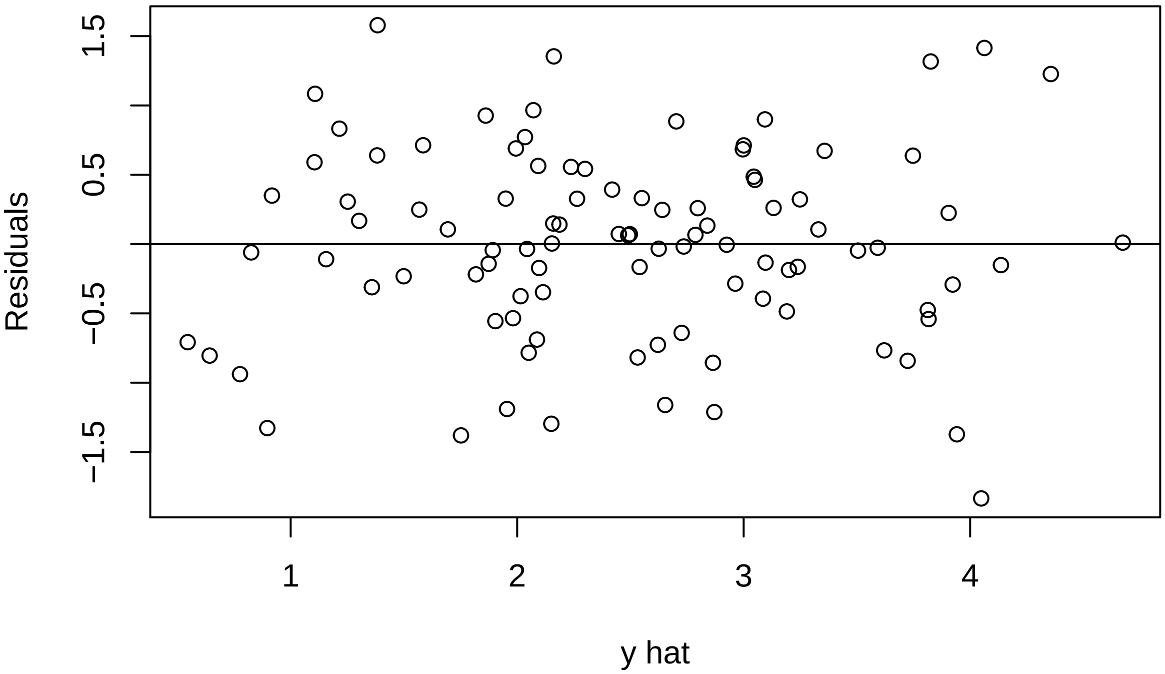
```
## [1] 2.00821
```

This interval does include the observed value.

e) It is often useful to make a scatter plot of the residuals, $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$, on the vertical axis versus the fitted values, $\hat{\mathbf{y}}$ on the horizontal axis. Make such a scatter plot for the model in Part b), and add a horizontal line at a vertical height 0. Hint: `abline(h=0)`. Is there any trend or pattern, or does the point cloud appear random without any systematic trend or curvature? Explain briefly.

Answer:

```
graph = plot(fit_2b$fitted.values, fit_2b$residuals,
             xlab = "y hat",
             ylab = "Residuals") +
  abline(h = 0)
```



There seems to be no clear trend or pattern between residuals and fitted values. The residuals seem to be fairly clustered around the residuals = 0 line regardless of the y hat value and there appears to be no extreme outliers.

Problem 2

This problem considers the `prostate` data in the library `faraway`. See `help(prostate)` for more information about the data set.

a) Fit a linear model with `lpsa` as the response and all the other variables as predictor variables. Display the model summary and identify all variables whose coefficient t values are statistically significant at the $\alpha = 0.05$ level. In each case, state what the null hypothesis is for the t test.

Answer:

```r
library(faraway)
fit_2a = lm(lpsa ~ ., data = prostate)
summary(fit_2a)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

The lcavol, lweight, and svi predictors are statistically significant at the $\alpha = 0.05$ significance level. The null hypothesis in each case is that the coefficient of a single predictor is 0 and that there is no linear association between this predictor and the lpsa response variable after adjusting for all other predictors.

b) Fit a reduced model with `lpsa` as the response but removing `lcp`, `gleason`, and `pgg45` as predictors in the model. Display the model summary, and provide both 90% and 95% confidence intervals for the coefficient of `lbph`. Do either or both of them include 0?

Answer:

```
fit_2b = lm(lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)
summary(fit_2b)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83505 -0.39396  0.00414  0.46336  1.57888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.95100    0.83175   1.143 0.255882
## lcavol       0.56561    0.07459   7.583 2.77e-11 ***
## lweight      0.42369    0.16687   2.539 0.012814 *
## age         -0.01489    0.01075  -1.385 0.169528
## lbph         0.11184    0.05805   1.927 0.057160 .
## svi          0.72095    0.20902   3.449 0.000854 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7073 on 91 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6245
## F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

```
confint(fit_2b, 'lbph', level = .95)
```

```
##             2.5 %    97.5 %
## lbph -0.003474551 0.2271544
```

```
confint(fit_2b, 'lbph', level = .90)
```

```
##             5 %      95 %
## lbph 0.01536969 0.2083102
```

Only the 95% confidence interval includes 0. This makes sense because the p-value of the t-test for the lbph predictor outputted in the summary is low enough to reject at the 90% confidence level, but not at the 95% confidence level.

c) Perform an F test comparing the model from Part b) as the null model, and the model from Part a) as the alternative model. Based on your calculations, does the test reject or accept the null hypothesis at the level $\alpha = 0.05$?

Answer:

```
anova(fit_2b, fit_2a)
```

```
## Analysis of Variance Table
##
## Model 1: lpsa ~ lcavol + lweight + age + lbph + svi
## Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     91 45.526
## 2     88 44.163  3    1.3625 0.905 0.4421
```

With a p-value of 0.4421, this test fails to reject the null hypothesis at the $\alpha = 0.05$ level. We accept the reduced model over the full model.

d) Here are the data for observation #32:

```
> prostate[32,]
      lcavol lweight  age    lbph   svi    lcp gleason pgg45    lpsa
32 0.1823216  6.1076  65 1.704748   0 -1.38629       6     0 2.00821
```

Pretending you did not know the value for `lpsa` for this observation, use your model from Part b) to compute a 95% prediction interval for observation 32 based on the predictors in the model. Does the interval include the value observed in the data?

Answer:

```
predict(fit_2b, newdata = prostate[32,], interval = "prediction", level = .95)
```

```
##          fit      lwr      upr
## 32 2.864524 1.244189 4.484858
```

```
prostate[32,]$lpsa
```

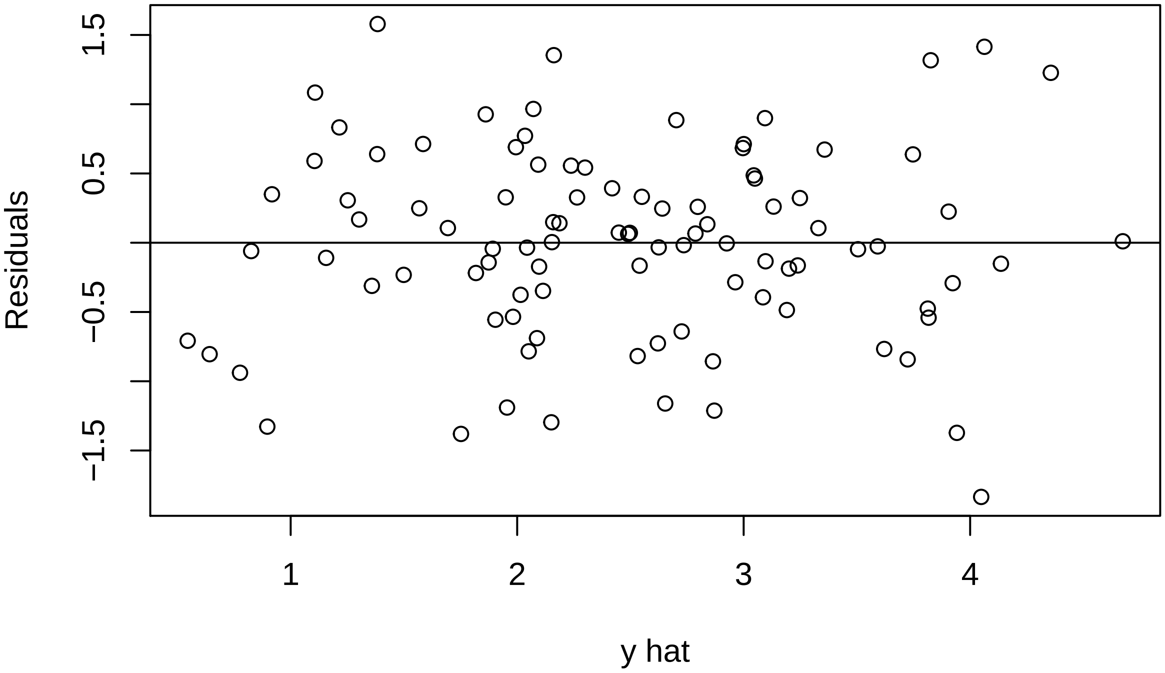
```
## [1] 2.00821
```

This interval does include the observed value.

e) It is often useful to make a scatter plot of the residuals, $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$, on the vertical axis versus the fitted values, $\hat{\mathbf{y}}$ on the horizontal axis. Make such a scatter plot for the model in Part b), and add a horizontal line at a vertical height 0. Hint: `abline(h=0)`. Is there any trend or pattern, or does the point cloud appear random without any systematic trend or curvature? Explain briefly.

Answer:

```
graph = plot(fit_2b$fitted.values, fit_2b$residuals,
             xlab = "y hat",
             ylab = "Residuals") +
  abline(h = 0)
```



There seems to be no clear trend or pattern between residuals and fitted values. The residuals seem to be fairly clustered around the residuals = 0 line regardless of the y hat value and there appears to be no extreme outliers.

Problem 3:

This problem refers to the `punting` data in the `faraway` library. The average distance punted and hang times of 10 punts of a football were measured for 13 volunteers. The left and right leg strength and flexibility were also recorded for each volunteer.

a) Fit a regression model with `Distance` as the response, and `RStr`, `LStr`, `RFlex` and `LFlex` as predictors (left and right leg strength, and left and right leg flexibility). Present a summary of the fitted model. Which if any predictors are significant at the 5% level?

Answer:

```
library("faraway")
model = lm(Distance~RFlex + LFlex + LStr + RStr, data = punting)
summary(model)
```

```
##
## Call:
## lm(formula = Distance ~ RFlex + LFlex + LStr + RStr, data = punting)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.941  -8.958  -4.441  13.523  17.016
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -79.6236    65.5935  -1.214    0.259
## RFlex         2.3745     1.4374   1.652    0.137
## LFlex        -0.5277     0.8255  -0.639    0.541
## LStr         -0.1862     0.5130  -0.363    0.726
## RStr          0.5116     0.4856   1.054    0.323
##
## Residual standard error: 16.33 on 8 degrees of freedom
## Multiple R-squared:  0.7365, Adjusted R-squared:  0.6047
## F-statistic:  5.59 on 4 and 8 DF,  p-value: 0.01902
```

As we can see from p value column in the summary of the model, none of the variables are significant at this level since they are all greater than 0.05.

b) Use an F-test to determine whether collectively these four predictors have any relationship with the response, i.e., test the (null) hypothesis that $\beta_{\texttt{RStr}} = \beta_{\texttt{LStr}} = \beta_{\texttt{RFlex}} = \beta_{\texttt{LFlex}} = 0$. (Here we are referring to the coefficient for predictor X_j in the model as β_{X_j}.) What do you conclude?

Answer:

```
modelreduce = lm(Distance~1 , data  = punting)
anova(modelreduce, model)
```

```
## Analysis of Variance Table
##
## Model 1: Distance ~ 1
## Model 2: Distance ~ RFlex + LFlex + LStr + RStr
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     12 8093.3
## 2      8 2132.6  4    5960.7 5.5899 0.01902 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p value is smaller than 0.05, which means we reject the null hypothesis that all the coefficients are 0.

c) Now we wish to test whether $\beta_{\texttt{RStr}} = \beta_{\texttt{LStr}}$ but not necessarily 0. Under the hypothesis that these two coefficients are equal write out the regression model formula and show that it is equivalent to replacing `RStr` and `LStr` in the model by the single variable `Str` = `RStr` + `LStr`.

Answer: If these two coefficients are equal ($\beta_{RStr} = \beta_{LStr} = \beta_{Str}$), the regression model could be written as:

$$\begin{aligned} y &= \beta_{RStr}x_{RStr} + \beta_{LStr}x_{LStr} + \beta_{RFlex}x_{RFlex} + \beta_{LFlex}x_{LFlex} + e \\ &= \beta_{Str}x_{RStr} + \beta_{Str}x_{LStr} + \beta_{RFlex}x_{RFlex} + \beta_{LFlex}x_{LFlex} + e \\ &= \beta_{Str}(x_{RStr} + x_{LStr}) + \beta_{RFlex}x_{RFlex} + \beta_{LFlex}x_{LFlex} + e \\ &= \beta_{Str}x_{Str} + \beta_{RFlex}x_{RFlex} + \beta_{LFlex}x_{LFlex} + e \end{aligned}$$

We can see it is equivalent to substitute these two variables with a single one by defining the new one “Str” as sum of “RStr” and “LStr”.

d) Use an F-test to test whether $\beta_{\texttt{RStr}} = \beta_{\texttt{LStr}}$. Note that the reduced model implied by this hypothesis entails replacing `RStr` and `LStr` in the `lm` model formula by `I(RStr+LStr)` (using the syntax of R).

Answer:

```r
modelred = lm(Distance ~ I(RStr + LStr) + RFlex + LFlex, data = punting)
anova(modelred, model)
```

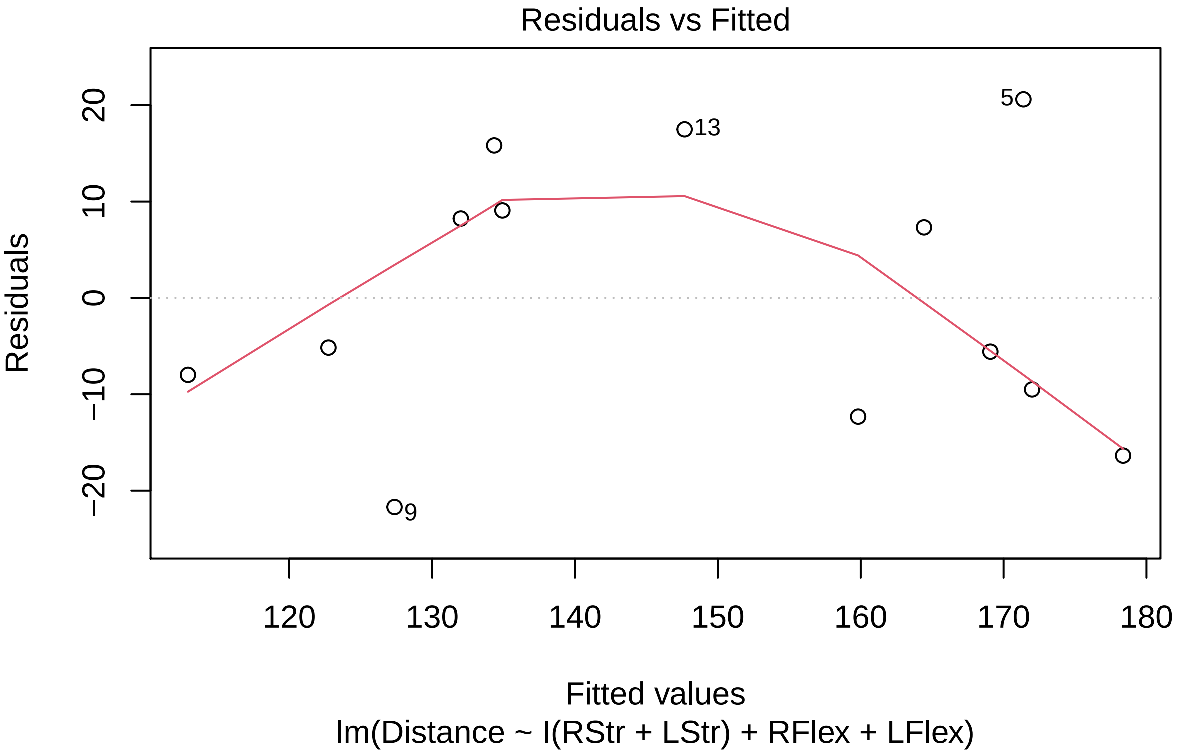
```
## Analysis of Variance Table
##
## Model 1: Distance ~ I(RStr + LStr) + RFlex + LFlex
## Model 2: Distance ~ RFlex + LFlex + LStr + RStr
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      9 2287.4
## 2      8 2132.6  1    154.72 0.5804  0.468
```

From the test result, p value is much larger than 0.05. The result shows the the null hypothesis is acceptable. Therefore, we can accept these two coefficients are equal.

e) Make a plot of residuals versus fitted values for the reduced model considered in Part d). Does the plot show any trend or pattern, or does it appear to be random noise? Explain briefly.

Answer:

```
plot(modelred, which = 1)
```



From the plot we can see there is an obvious trend. But the points seem to be evenly spread around the trend so the noise appears to be random.

STAT 425 Assignment 3

Due Sunday, March 7, 11:59pm. Submit through Moodle.

Name: (insert your name here)

Netid: (insert)

Submit your computational work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown.

Problem 1

The `fat` data in the `faraway` library contains age, weight, height, and various body circumference measurements for 252 men. The variables `brozek` and `siri` correspond to two different density related equations for percent body fat.

a) Fit a linear model to predict the `siri` percentage body fat from `age`, `weight`, `height`, `neck`, `chest`, `abdom`, `hip`, `thigh`, `knee`, `ankle`, `biceps`, `forearm`, and `wrist`. Display a model summary and state which of the variable coefficients are statistically significant at the 0.05 level.

Answer: We first fit the model as below.

```
# load package
require('faraway')
```

```
## Loading required package: faraway
```

```
lmod <- lm(siri ~ age + weight + height + neck + chest + abdom +
             hip + thigh + knee + ankle + biceps + forearm +
             wrist, data = fat)
summary(lmod)
```

```
##
## Call:
## lm(formula = siri ~ age + weight + height + neck + chest + abdom +
##     hip + thigh + knee + ankle + biceps + forearm + wrist, data = fat)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -11.1687  -2.8639  -0.1014   3.2085  10.0068 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -18.18849   17.34857  -1.048  0.29551    
## age           0.06208    0.03235   1.919  0.05618 .  
## weight       -0.08844    0.05353  -1.652  0.09978 .  
## height       -0.06959    0.09601  -0.725  0.46925    
## neck         -0.47060    0.23247  -2.024  0.04405 *  
## chest        -0.02386    0.09915  -0.241  0.81000    
## abdom         0.95477    0.08645  11.044  < 2e-16 ***
## hip          -0.20754    0.14591  -1.422  0.15622    
## thigh         0.23610    0.14436   1.636  0.10326    
## knee          0.01528    0.24198   0.063  0.94970    
## ankle         0.17400    0.22147   0.786  0.43285    
## biceps        0.18160    0.17113   1.061  0.28966    
## forearm       0.45202    0.19913   2.270  0.02410 *  
## wrist        -1.62064    0.53495  -3.030  0.00272 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.305 on 238 degrees of freedom
## Multiple R-squared:  0.749,  Adjusted R-squared:  0.7353 
## F-statistic: 54.65 on 13 and 238 DF,  p-value: < 2.2e-16
```

As we can see from the summary table, under a significance level of 5%, the coefficients of the circumferences of neck, abdom, forearm, and wrist are statistically significant.

b) Perform a model comparison F test to determine whether a reduced model using only the variables `age`, `weight`, `height`, `abdom`, `forearm`, and `wrist` to predict `siri` is appropriate. Is the null hypothesis rejected at level 0.05?

Answer: We can fit the reduced model, and perform the ANOVA F-test as below.

```r
submod <- lm(siri ~ age + weight + height + abdom + forearm +
               wrist, data = fat)
anova(submod, lmod)
```

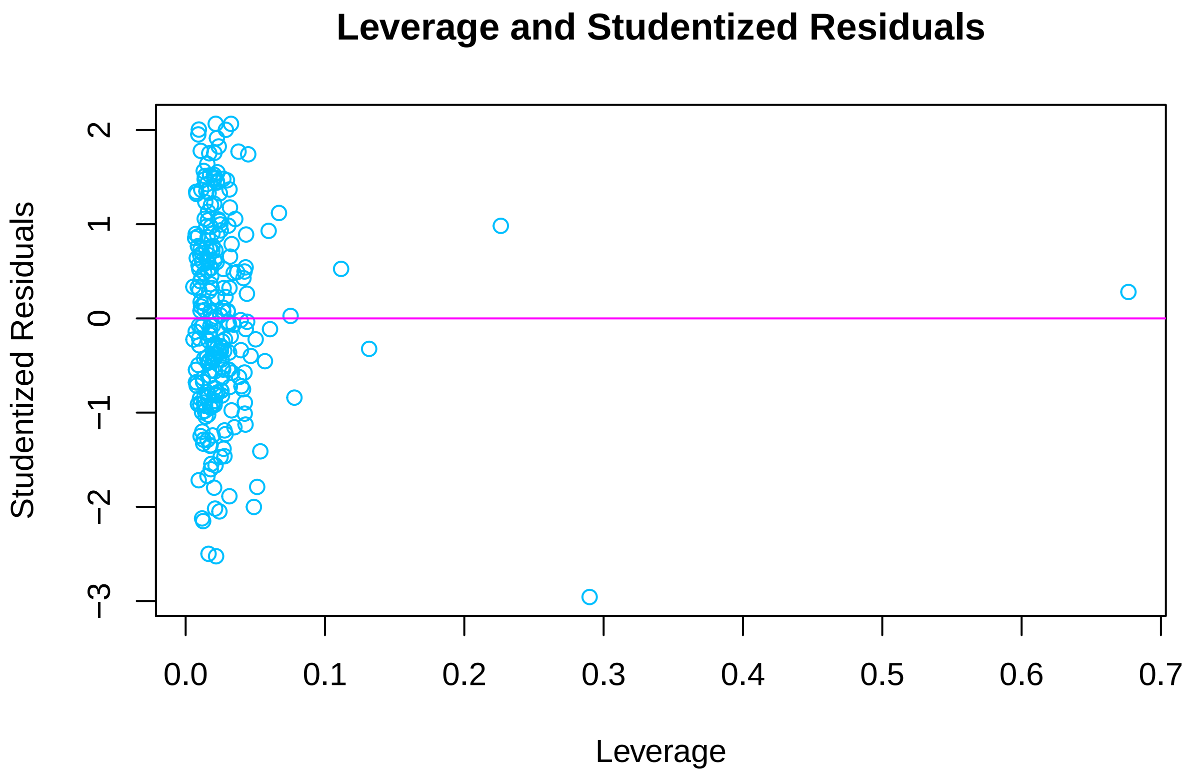
```
## Analysis of Variance Table
## 
## Model 1: siri ~ age + weight + height + abdom + forearm + wrist
## Model 2: siri ~ age + weight + height + neck + chest + abdom + hip + thigh + 
##     knee + ankle + biceps + forearm + wrist
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    245 4601.8
## 2    238 4411.4  7    190.34 1.467 0.1798
```

Under a significance level of 5%, with p-value 0.1798, we fail to reject the null hypothesis.

c) Now consider the reduced model from Part b). Plot studentized residuals on the y-axis versus leverage (diagonal of hat matrix) on the x-axis for this model. Include a horizontal line at height 0 for reference. Recall from our notes that the function `influence` provides diagonals of the hat matrix as influence(ModelObject)$hat. Are there any high leverage observations, and if so do they appear to be response outliers, or are they well fit by the model?

Answer: We first plot

```r
plot(x = influence(submod)$hat, y = rstudent(submod),
     xlab = 'Leverage', ylab = 'Studentized Residuals',
     col = 'deepskyblue',
     main = 'Leverage and Studentized Residuals')
abline(h = 0, col = 'magenta')
```



Using a rule of thumb of $2p/n$, where $n = 252$ and $p = 7$, we can find observations with high leverage as below.

```r
p <- ncol(model.matrix(submod))
n <- nrow(fat)
(hlobs <- which(influence(submod)$hat > 2 * p / n))
```

```
##  36  39  41  42 159 175 205 206 216 226 252
##  36  39  41  42 159 175 205 206 216 226 252
```

We can check whether they are outliers using Bonferroni correction as

```
m <- n
cv <- qt(.05 / (2 * m), df = df.residual(submod))
which(abs(rstudent(submod)[hlobs]) > abs(cv))
```

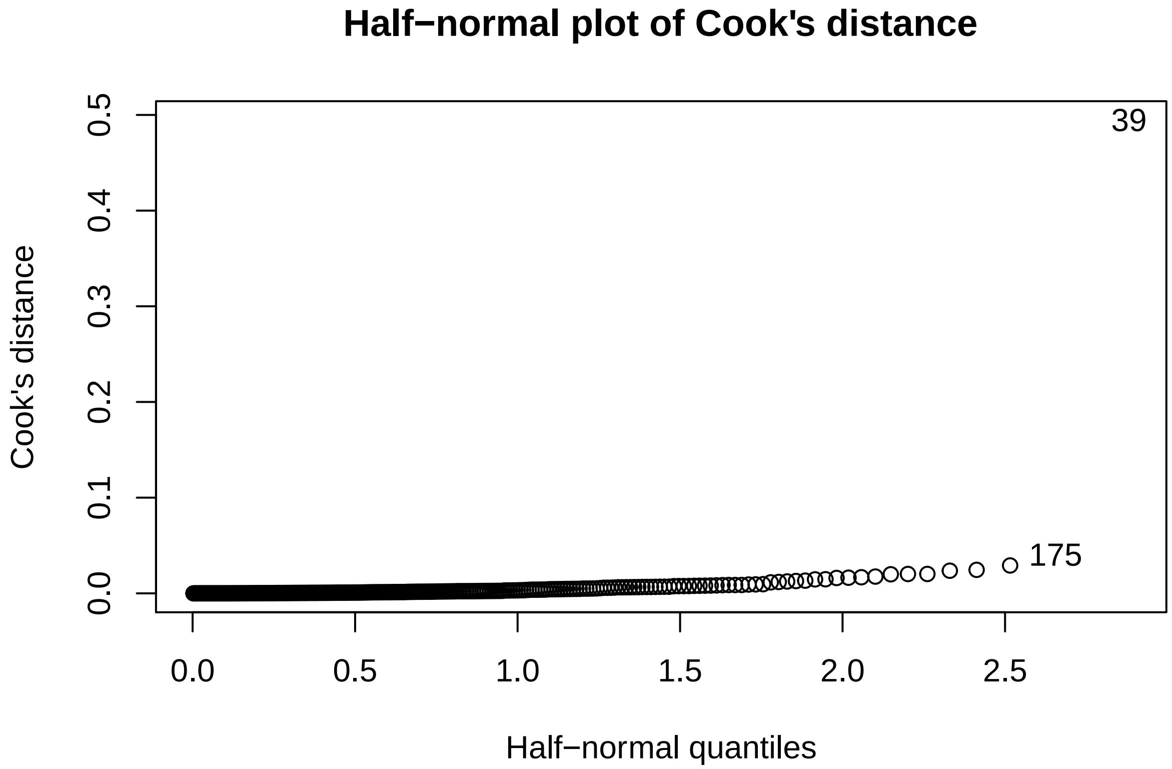
```
## named integer(0)
```

None of the observations is rejected as an outlier after Bonferroni adjustment for the sample size.

d) For the reduced model in parts b) and c), use a half-normal plot or sort function to display which observation has the largest Cook's Distance.

Answer: We can find the observation that has the largest Cook's distance as below.

```
cook <- cooks.distance(submod)
halfnorm(cook,  ylab = "Cook's distance",
         main = "Half-normal plot of Cook's distance")
```



Thus, the 39th observation has the largest Cook's distance.

e) For the same reduced model as in Part d), show the model summaries with and without the observation identified by Cook's Distance in Part d). Indicate which, if any, of the variables have significant coefficient t-tests in one but not the other of these two fitted models, at a significance level of 0.05.

Answer: We can check this result with

```
submod2 <- lm(siri ~ age + weight + height + abdom + forearm +
                wrist, data = fat[-39, ])
tab <- round(cbind(coef(summary(submod))[, 4],
      coef(summary(submod2))[, 4], NA), 4)
tab[, 3] <- ifelse((tab[, 1] >= 0.05 & tab[, 2] >= 0.05 |
                     tab[, 1] < 0.05 & tab[, 2] < 0.05) == TRUE,
                   yes = 0, no = 1)
colnames(tab) <- c('With', 'Without', 'Inconsistent')
tab
```

```
##               With Without Inconsistent
## (Intercept) 0.0094  0.0575            1
## age         0.1541  0.0927            0
## weight      0.0020  0.0704            1
## height      0.3275  0.1530            0
## abdom       0.0000  0.0000            0
## forearm     0.0069  0.1051            1
## wrist       0.0005  0.0002            0
```

As we can see, `weight` and `forearm` have significant coefficient *t*-tests in one but not the other of these two fitted models, at a significance level of 5%.

Problem 2

The `cheddar` data from the `faraway` library has data on samples of cheddar cheese. For each of 30 samples, a subjective `taste` score is given along with checmical composition measurements labeled `Acetic`, `H2S`, and `Lactic`.

a) Fit a multiple linear regression model for predicting `taste` from the three chemical composition measurements. Which of the variables have statistically significant coefficients at level 0.05?

Answer:

```
library(faraway)
fit = lm(taste ~ Acetic + H2S + Lactic, data = cheddar)
summary(fit)
```

```
##
## Call:
## lm(formula = taste ~ Acetic + H2S + Lactic, data = cheddar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
```

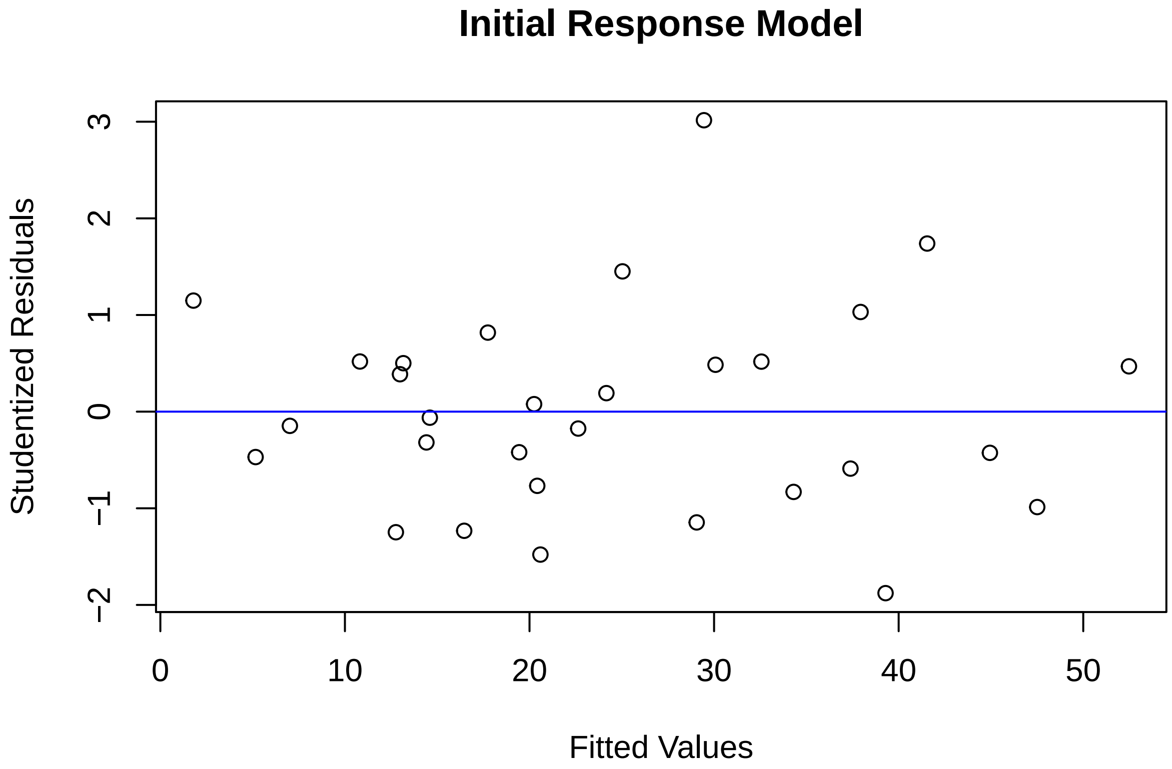
```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## Acetic        0.3277     4.4598   0.073  0.94198
## H2S           3.9118     1.2484   3.133  0.00425 **
## Lactic       19.6705     8.6291   2.280  0.03108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.13 on 26 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6116
## F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

H2S and Lactic have statistically significant coefficients at the the 0.05 significance level.

b) Plot studentized residuals versus fitted values for the model in a), including a horizontal line at height 0 for reference. Is there any evidence of outliers, heteroscedasticity or curvature in the plot? Explain briefly.

Answer:

```r
p = plot(fit$fitted.values, rstudent(fit),
     xlab = "Fitted Values",
     ylab = "Studentized Residuals",
     main = "Initial Response Model") +
  abline(h = 0, col = "blue")
```



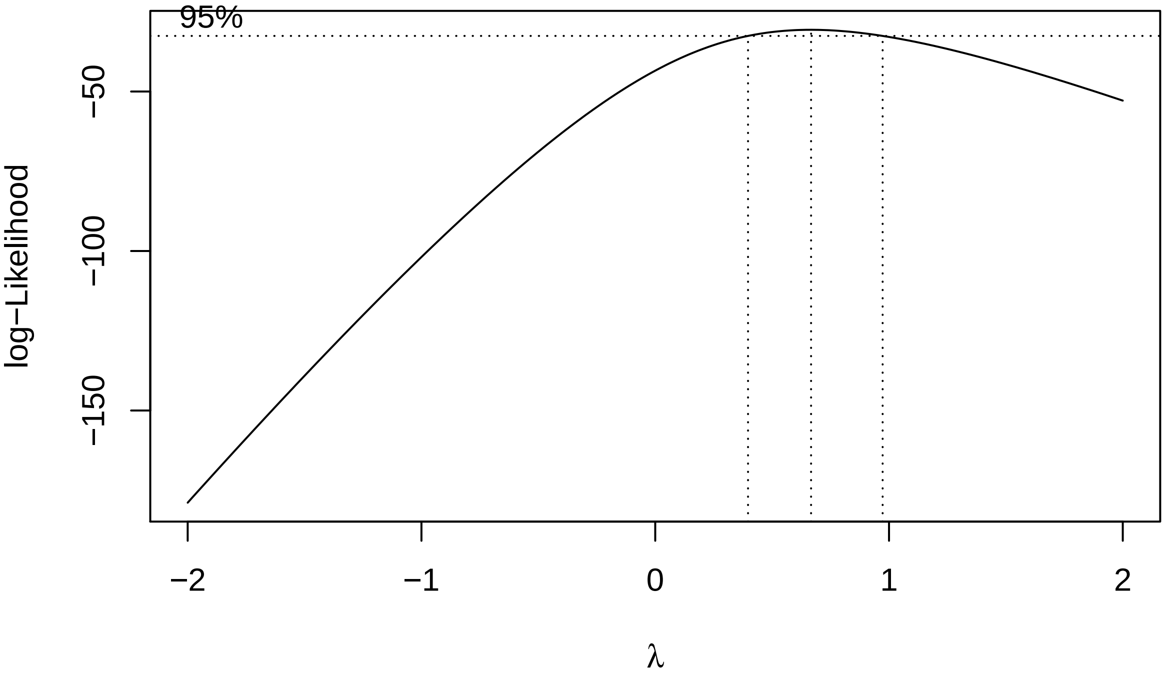
There seems to be a point with a studentized residual value of about 3 that may be an outlier.

The studentized residuals seem to have slightly higher variance with greater fitted values, so that may be evidence for some slight heteroscedasticity. Other than that, the points seem to be distributed fairly randomly, so there is no evidence for a clear curve in the plot.

c) Use the Box-Cox method to find an optimal transformation of the response. What is the optimal power $\hat{\lambda}$ of the response, based on the log-likelihood?

Answer:

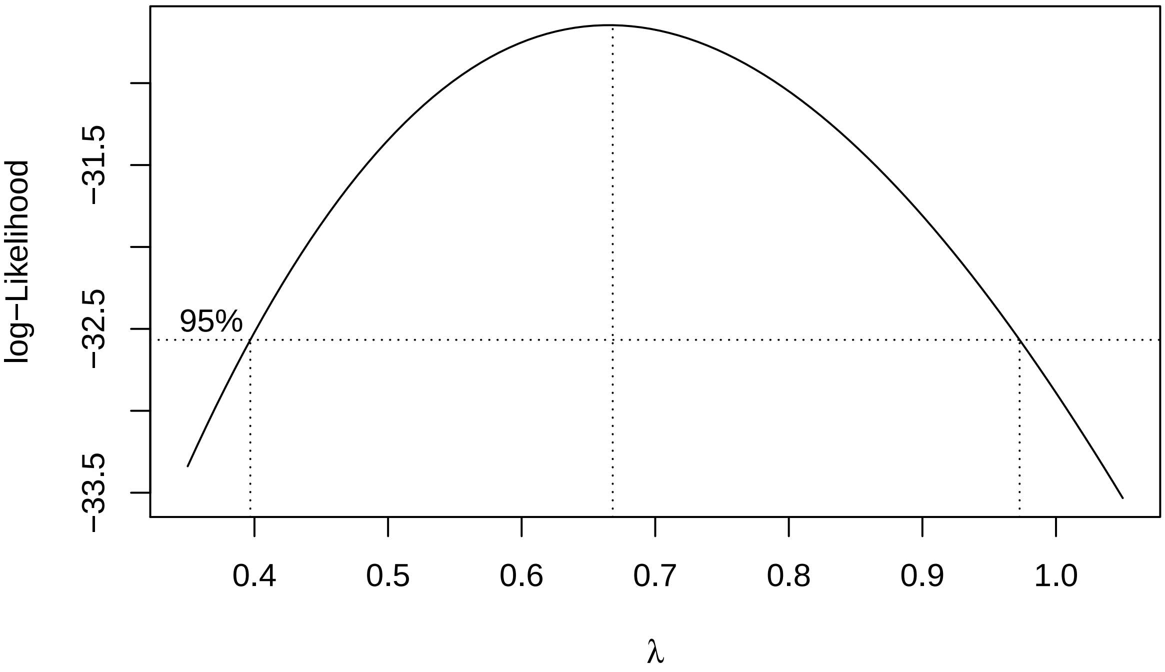
```
library(MASS)
trans = boxcox(fit, lambda=seq(-2, 2, length=400))
```



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

```
## [1] 0.6666667
```

```
boxcox(fit,plotit=T,lambda=seq(0.35,1.05,by=0.1))
```



The optimal lambda value for the response is .6667 or 2/3.

d) Does the 95% confidence interval for the Box-Cox transformation parameter λ include $\lambda = 1$? What is your interpretation of this?

Answer:

```
tmp=trans$x[trans$y > max(trans$y) - qchisq(0.95, 1)/2]
range(tmp)
```

```
## [1] 0.4060150 0.9674185
```

This confidence interval doesn't include the value of lambda = 1, so an appropriate transformation is encouraged.

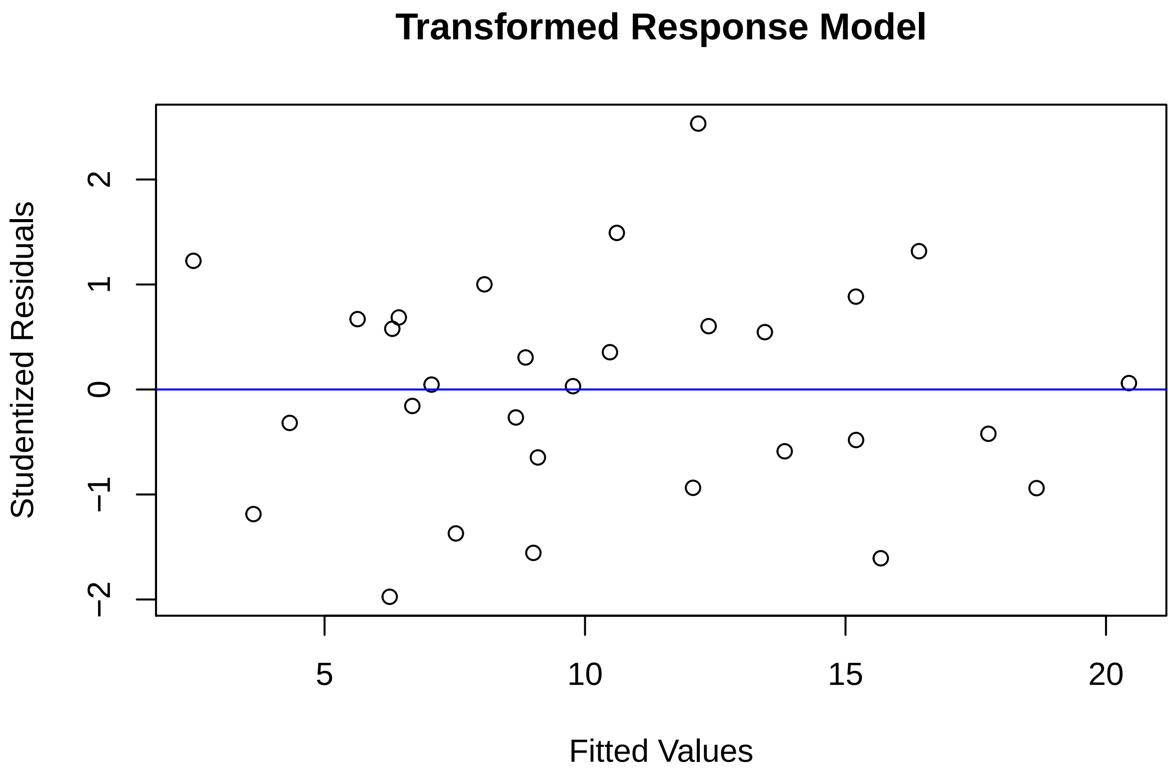
e) Use the optimal transformation found in c) to refit the data with the transformed response depending on the three chemical composition measurements. Plot studentized residuals versus fitted values for this transformed response model. Compare the graph qualitatively to the graph in Part b). Note: you can include a transformation of a variable in a formula by enclosing the transformation inside the function `I( )`, as in lm(I(y^3)~x, . . .).

Answer:

```
lambda = 2/3
fit_2 = lm(I((taste^(lambda) - 1)/lambda) ~ Acetic + H2S + Lactic, data = cheddar)

p2 = plot(fit_2$fitted.values, rstudent(fit_2),
     xlab = "Fitted Values",
     ylab = "Studentized Residuals",
     main = "Transformed Response Model") +
```

```r
  abline(h = 0, col = "blue")
```



After the transformation, the single point that may have been an outlier seems less extreme. The variance across studentized residuals is fairly even regardless of the fitted values. This would suggest the transformed model has reduced heteroscedasticity when compared to the initial model. The curvature appears to be relatively the same.

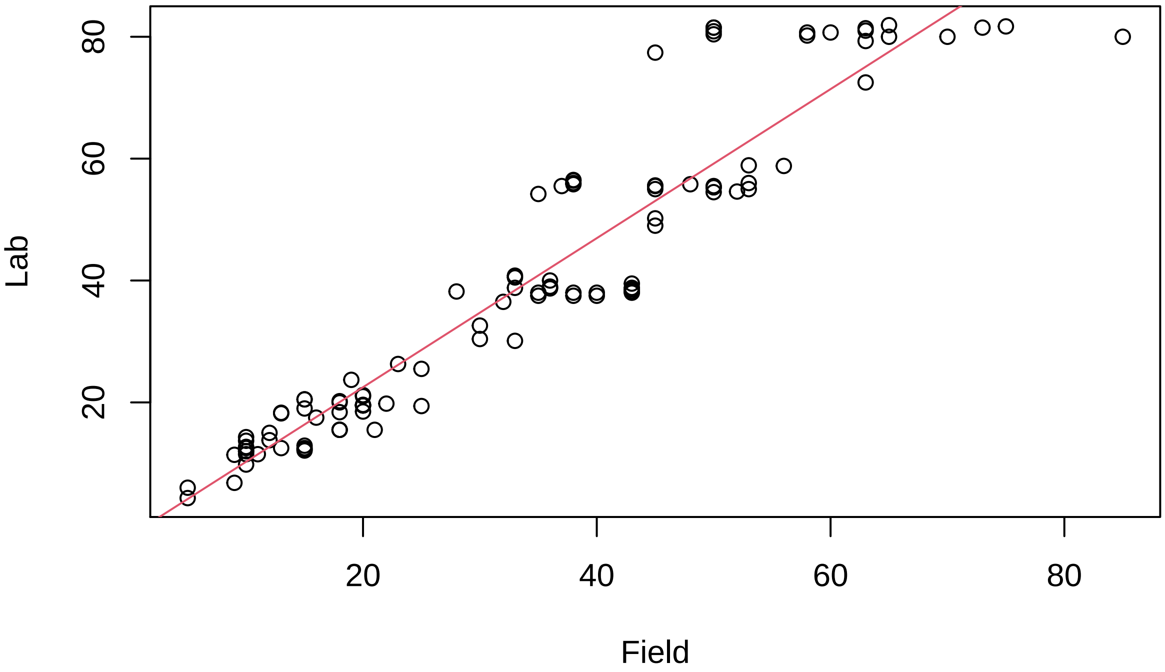
Problem 3:

The `pipeline` data in the `faraway` library consist of ultrasonic measurements of defects in the Alaska pipeline in the `Lab` and in the `Field`, i.e., on site.

a) Make a scatter plot of the `Lab` measurements versus the `Field` measurements, including the least squares regression line on the plot.

Answer:

```r
library("faraway")
model = lm(Lab ~ Field, data = pipeline)
plot(pipeline$Field, pipeline$Lab, xlab = "Field", ylab = "Lab")
abline(model$coefficients, col = 2)
```



b) Fit the linear regression of `Lab` on `Field` and show the model summary. How strong is the linear correlation between the lab and field measurements?

Answer:

```
summary(model)
```

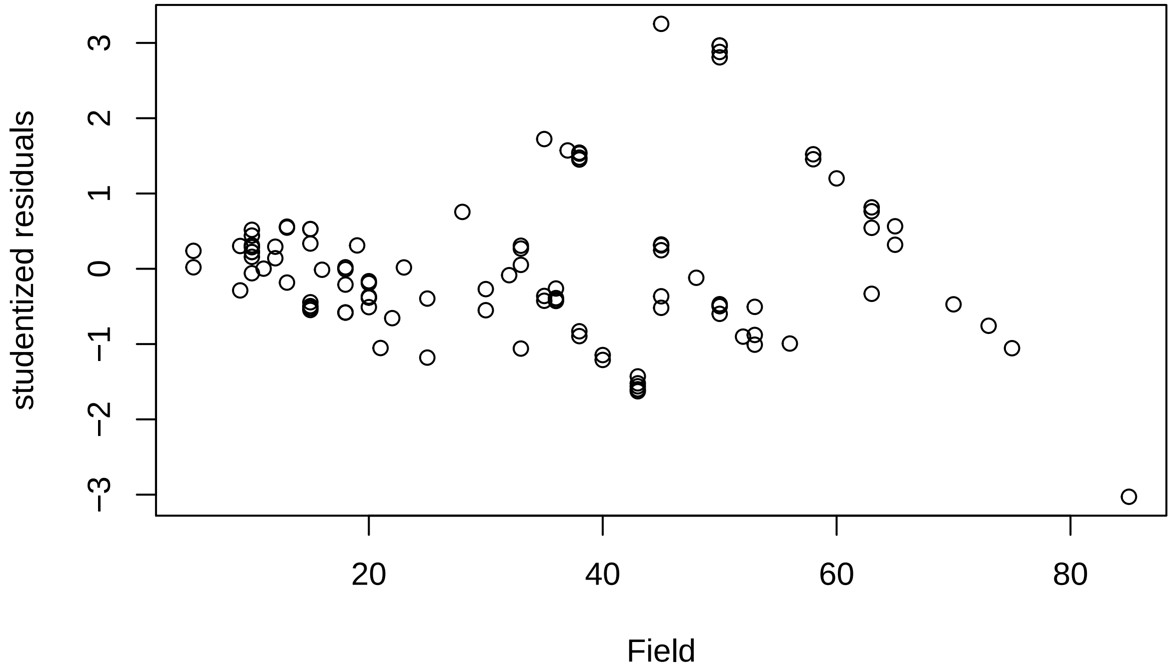
```
##
## Call:
## lm(formula = Lab ~ Field, data = pipeline)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.985  -4.072  -1.431   2.504  24.334
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.96750    1.57479  -1.249    0.214
## Field        1.22297    0.04107  29.778   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.865 on 105 degrees of freedom
## Multiple R-squared:  0.8941, Adjusted R-squared:  0.8931
## F-statistic: 886.7 on 1 and 105 DF,  p-value: < 2.2e-16
```

The multiple R^2 is 0.8941 according the model summary. We can see the linear correlation is pretty strong between the lab and field.

c) Plot studentized residuals versus `Field` for the model in b). Does the graph show any evidence of heteroscedascticity or curvature? Describe briefly.

Answer:

```
plot(pipeline$Field, rstudent(model),
     xlab = "Field",
     ylab = "studentized residuals")
```



We somehow can see the trend here but apparently the variance of error is not constant according to the plot. This suggests heteroscedasctic error variance.

d) Fit a linear model regressing the log of `Lab` on the log of `Field`. Show the model summary and compare the linear correlation with that of the model in b).

Answer:

```
modellog = lm(I(log(Lab))~I(log(Field)), data = pipeline)
summary(modellog)
```

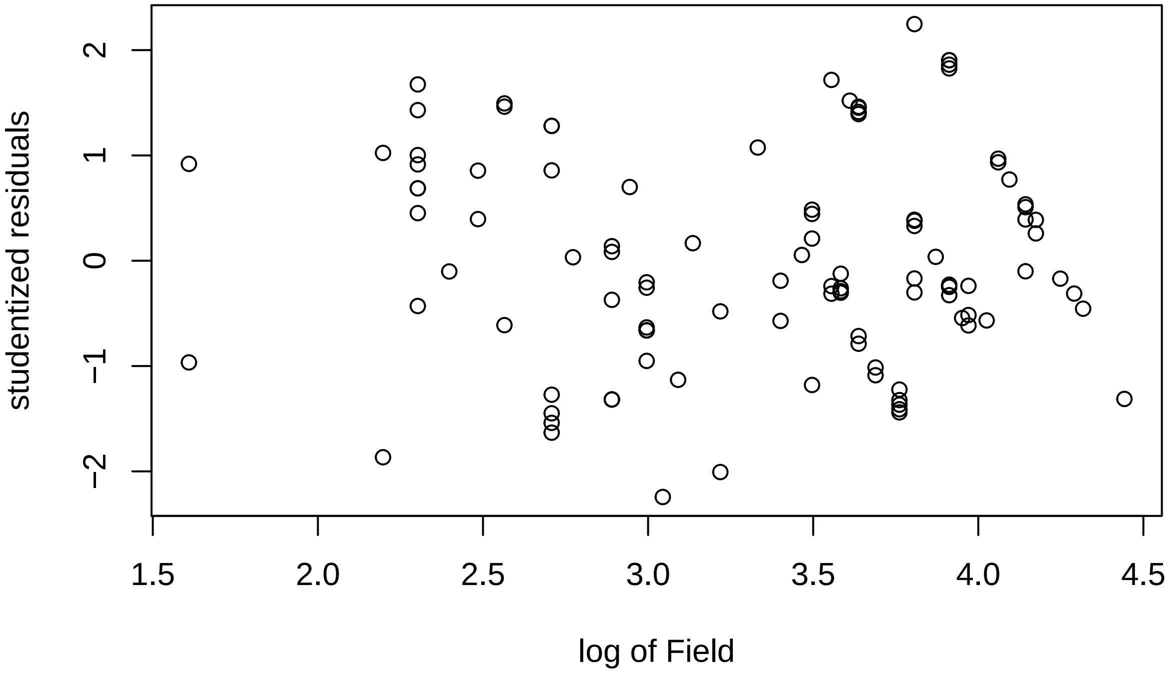
```
##
## Call:
## lm(formula = I(log(Lab)) ~ I(log(Field)), data = pipeline)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40212 -0.11853 -0.03092  0.13424  0.40209
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.06849    0.09305  -0.736    0.463
## I(log(Field))  1.05483    0.02743  38.457   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1837 on 105 degrees of freedom
## Multiple R-squared:  0.9337, Adjusted R-squared:  0.9331
## F-statistic:  1479 on 1 and 105 DF,  p-value: < 2.2e-16
```

The multiple R^2 value 0.9337 is larger than 0.8941. We can see that the linear correlation here is stronger.

e) For the model in d), plot studentized residuals versus the log of `Field`. Does this graph show any evidence of heteroscedascticity or curvature? Describe briefly.

Answer:

```
plot(log(pipeline$Field), rstudent(modellog),
     xlab = "log of Field",
     ylab = "studentized residuals")
```



The graph looks much better now. No obvious curvature is shown and the variance of error seems to be consistent.

STAT 425 Assignment 4

Due Tuesday, March 22, 11:59pm. Submit through Moodle.

Name: (insert your name here)

Netid: (insert)

Submit your computational work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown.

Most relevant class notes: 4.1.Collinearity, 4.2.GLS, 4.3.TestFit, 5.1.Polynomial

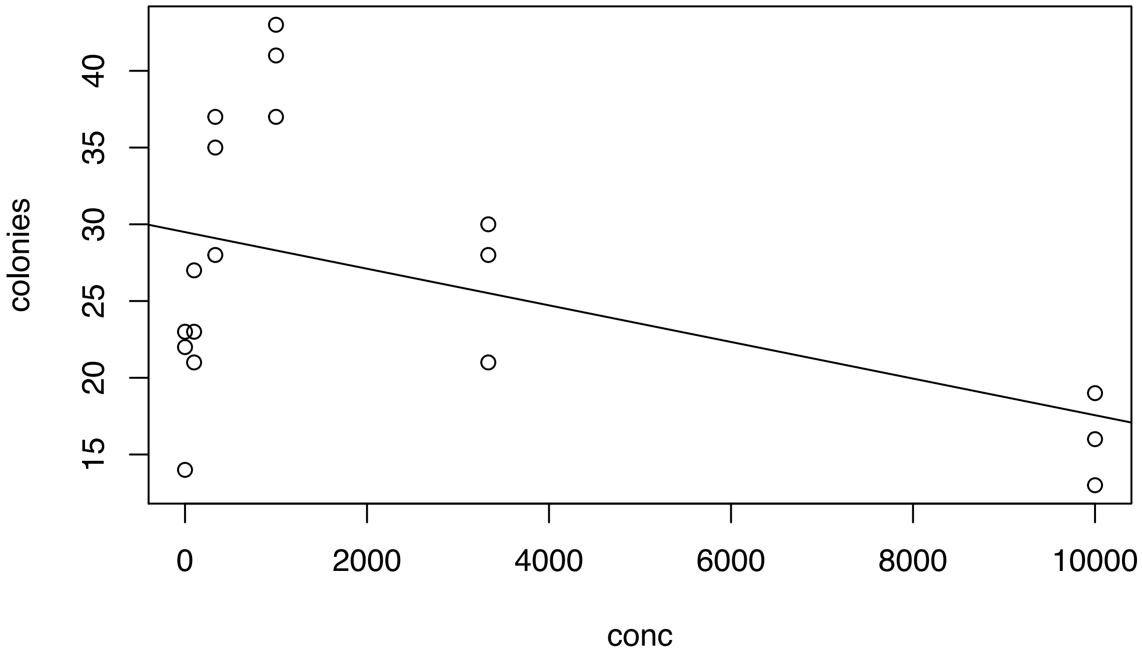
Problem 1

A study was conducted to see if a certain food dye (Acid Red 118) affected mutation rates in salmonella bacteria. The data are in the included file, "reddye.csv." The variables are concentration of red dye on the plate (`conc`), and number of mutation colonies developed on the plate (`colonies`).

a) Read the data into an R data frame, and make a scatterplot with `colonies` as the response and `conc` as the predictor. Include the least squares line on the graph. What does the slope of the LS line suggest about the relationship between `conc` and `colonies`? Does the line seem to fit the data?

Answer:

```r
data = read.csv("~/Desktop/STAT 425 TA spring/HW/hw4/reddye.csv")
attach(data)
plot_a = plot(conc, colonies) +
  abline(lm(colonies ~ conc))
```



The slope of the least squares line suggests that the relationship between conc and colonies is negative. The line seems to the fit the data fairly well, although the variance in colonies is higher when conc is lower.

b) Test the fit of the linear model in part a) versus the more general alternative model where conc is treated as a factor variable. Recall from class: if `conc` is treated as a factor variable, then we assume only that y has a different mean for each value of `conc` and nothing more. What do you conclude from the results?

Answer:

```r
fit_a = lm(colonies ~ conc)
fit_b = lm(colonies ~ factor(conc))
anova(fit_a, fit_b)
```

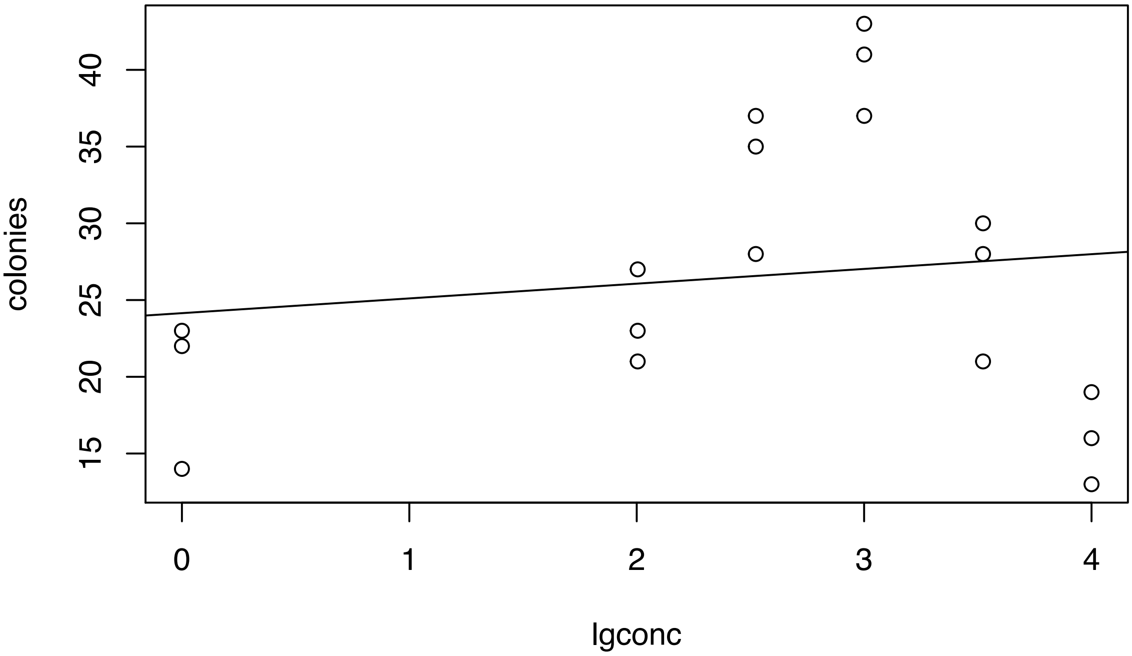
```
## Analysis of Variance Table
##
## Model 1: colonies ~ conc
## Model 2: colonies ~ factor(conc)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     16 1078.32
## 2     12  193.33  4    884.99 13.732 0.0001968 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p-value of 0.0001968, we can reject the null under any reasonable alpha and conclude that the model that treats conc as a factor variable is more suitable model.

c) Since the concentrations range over several orders of magnitude, a logarithmic transformation of `conc` might help. There is a problem, though, because the zero concentration would transform to $-\infty$. Instead, consider the constructed variable `lgconc=log10(1+conc)`. Make a scatterplot of `colonies` versus `lgconc`, including the least squares line for the corresponding linear regression model. What does the slope of the LS line suggest about the relationship between `lgconc` and `colonies`? Does the line seem to fit the data?

Answer:

```
lgconc = log10(1 + conc)
plot_c = plot(lgconc, colonies) +
  abline(lm(colonies ~ lgconc))
```



The slope of the least squares regression line suggests that the relationship between lgconc and colonies is fairly positive. The line does not seem to fit the data as the data takes on an almost parabolic shape rather than a linear one.

d) Test the fit of the model in c) by comparing it to the more general model that treats `lgconc` as a factor variable.

Answer:

```
fit_c = lm(colonies ~ lgconc)
fit_d = lm(colonies ~ factor(lgconc))
anova(fit_c, fit_d)
```

```
## Analysis of Variance Table
##
## Model 1: colonies ~ lgconc
## Model 2: colonies ~ factor(lgconc)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     16 1374.63
## 2     12  193.33  4    1181.3 18.331 4.765e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p-value of 4.765e-05, we have enough evidence to reject the null and assume that there is a lack of fit in the model in part c.

e) Obtain the R-square values and model F test p-values for three models:

$$\text{colonies} = \beta_0 + \beta_1 \texttt{conc} + \texttt{error}$$
$$\text{colonies} = \beta_0 + \beta_1 \texttt{lgconc} + \texttt{error}$$
$$\text{colonies} = \beta_0 + \beta_1 \texttt{lgconc} + \beta_2 \texttt{lgconc}^2 + \texttt{error}$$

Based on the results, which of these three models seems most reasonable and why? **Answer:**

```
fit_e1 = lm(colonies ~ conc)
fit_e2 = lm(colonies ~ lgconc)
fit_e3 = lm(colonies ~ lgconc + I(lgconc^2))

summary(fit_e1)
```

```
##
## Call:
## lm(formula = colonies ~ conc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.492  -5.920  -1.327   5.550  14.701
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 29.4919432  2.3530833  12.533 1.09e-09 ***
## conc        -0.0011932  0.0005441  -2.193   0.0434 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.209 on 16 degrees of freedom
## Multiple R-squared:  0.2311, Adjusted R-squared:  0.1831
## F-statistic: 4.809 on 1 and 16 DF,  p-value: 0.04343
```

```
summary(fit_e2)
```

```
##
## Call:
```

```
## lm(formula = colonies ~ lgconc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9885  -6.1654  -0.3378   6.9399  15.9719
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.1454     4.7664   5.066 0.000115 ***
## lgconc        0.9608     1.6887   0.569 0.577289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.269 on 16 degrees of freedom
## Multiple R-squared:  0.01983,    Adjusted R-squared:  -0.04143
## F-statistic: 0.3237 on 1 and 16 DF,  p-value: 0.5773
```

```
summary(fit_e3)
```

```
##
## Call:
## lm(formula = colonies ~ lgconc + I(lgconc^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7746  -4.7181   0.3789   4.6333  13.0402
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   18.217      4.284   4.253 0.000695 ***
## lgconc        14.002      4.405   3.179 0.006230 **
## I(lgconc^2)   -3.362      1.080  -3.113 0.007133 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.462 on 15 degrees of freedom
## Multiple R-squared:  0.4045, Adjusted R-squared:  0.3251
## F-statistic: 5.094 on 2 and 15 DF,  p-value: 0.0205
```

```
#Model 1: Multiple R squared value is .2311 and p-value is .04343
#Model 2: Multiple R squared value is .01983 and p-value is .5773
#Model 3: Multiple R squared value is .4045 and p-value is .0205
```

The third model seems to be the best fit out of the three because it has the lowest p-value at .0205 and the highest multiple R-squared value at .4045. A lower p-value suggests stronger evidence for a significant model and a higher multiple R-squared value means that the model's fitted values more closely resemble the response.

Problem 2

The `aatemp` data in the `faraway` library comes from the U.S. Historical Climatological Network. The data report annual mean temperatures in Ann Arbor Michigan for roughly 150 years.

a) Fit a linear trend model to temperature as a function of year and display the model summary. Does there appear to be a trend?

Answer:

```
library("faraway")
model = lm(temp ~ year, data = aatemp)
summary(model)
```

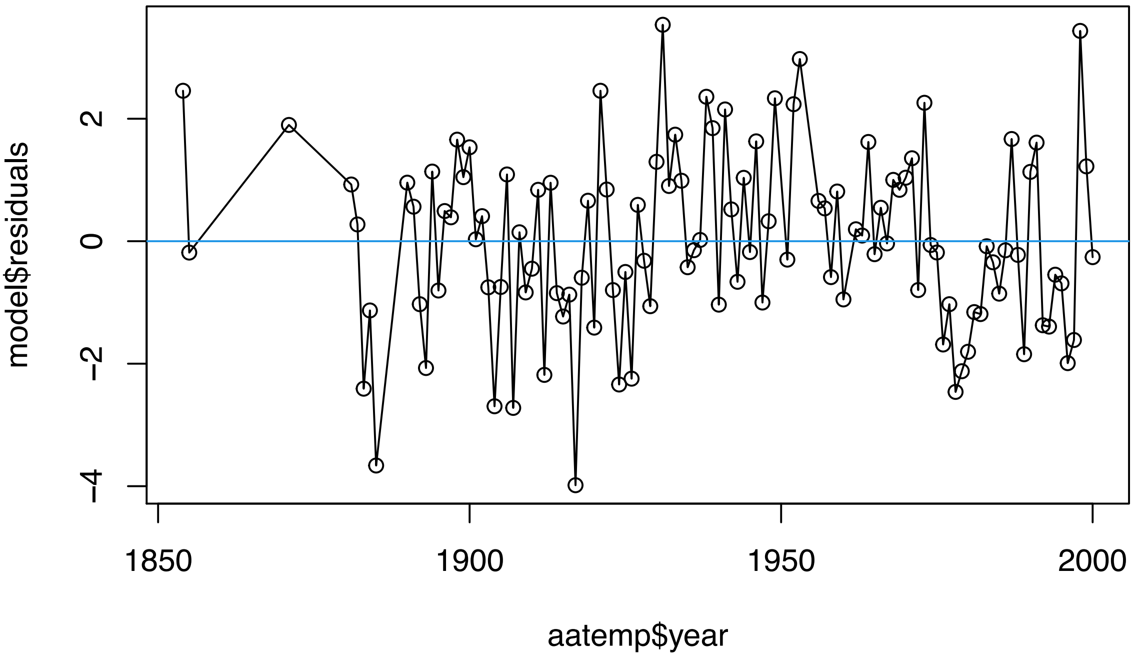
```
##
## Call:
## lm(formula = temp ~ year, data = aatemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9843 -0.9113 -0.0820  0.9946  3.5343
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.005510   7.310781   3.284  0.00136 **
## year         0.012237   0.003768   3.247  0.00153 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.466 on 113 degrees of freedom
## Multiple R-squared:  0.08536,    Adjusted R-squared:  0.07727
## F-statistic: 10.55 on 1 and 113 DF,  p-value: 0.001533
```

As we can see from the summary, p value for the slope coefficient is small, which implies there appears to be a trend.

b) For the model in a), plot residuals versus year, connecting the dots in the plot (`type='o'`), and adding a horizontal reference line at 0. Is there any evidence of serial correlation in the graph?

Answer:

```
plot(aatemp$year, model$residuals, type = 'o')
abline(h=0, col =4)
```



From the graph we can see the evidence of serial correction.

c) Based on the model in a), test for serial correlation between successive years using the Durbin-Watson test, and state your conclusion.

Answer:

```r
library("lmtest")
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
dwtest(model)
```

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 1.6177, p-value = 0.01524
## alternative hypothesis: true autocorrelation is greater than 0
```

Since the p value is smaller than 0.05, we may reject the null hypothesis and conclude that there exists serial correction.

d) Using the `gls` function from the `nlme` library, fit a linear trend model using the AR1 form of correlation between years. Display the model summary. Does the trend line change much? How much correlation is there, based on the estimated AR1 correlation parameter?

Answer:

```r
library("nlme")
armodel = gls(temp ~ year, correlation = corAR1(form = ~ year),
              data = aatemp)
summary(armodel)
```

```
## Generalized least squares fit by REML
##   Model: temp ~ year
##   Data: aatemp
##        AIC     BIC    logLik
##   426.5694 437.479 -209.2847
##
## Correlation Structure: ARMA(1,0)
##  Formula: ~year
##  Parameter estimate(s):
##      Phi1
## 0.2303887
##
## Coefficients:
##                Value Std.Error  t-value p-value
## (Intercept) 25.18407  8.971864 2.807006  0.0059
## year         0.01164  0.004626 2.516015  0.0133
##
##  Correlation:
##      (Intr)
## year -1
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.7230803 -0.6321970 -0.0520135  0.6645795  2.3775123
##
## Residual standard error: 1.475718
## Degrees of freedom: 115 total; 113 residual
```

The trend line does not seem changing much by checking the intercept and slope coefficients. Based on the model summary we can see the estimated AR1 correction is 0.2303887.

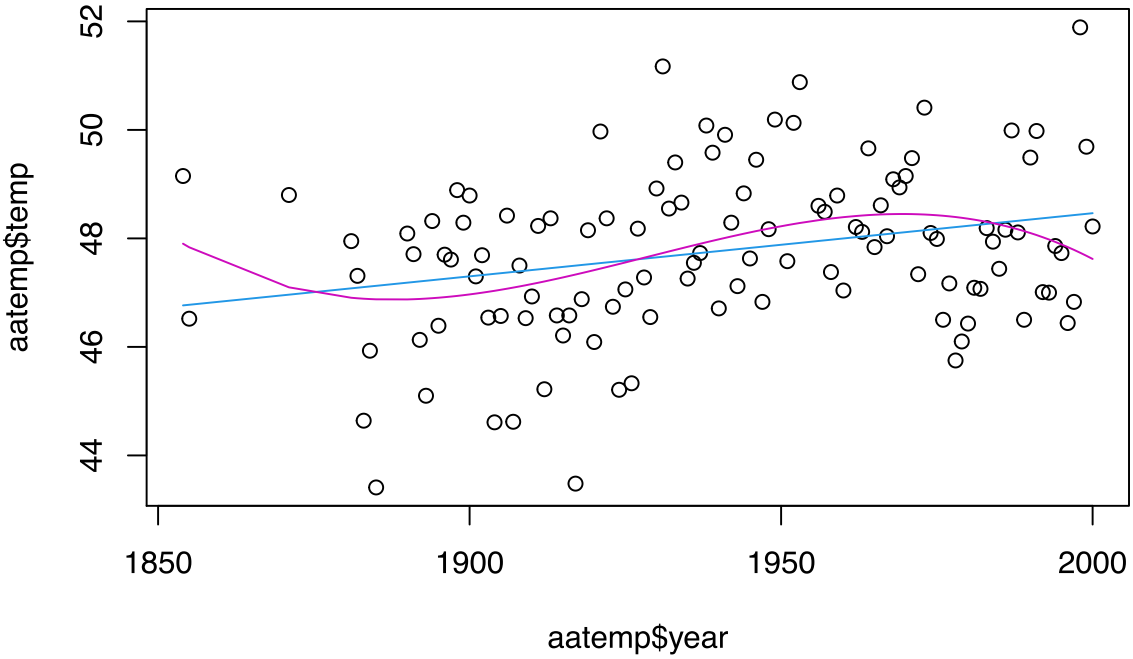
e) Again using the `gls` function, fit a cubic model (third order polynomial) for temperature as a function of year, with the AR1 form of correlation. Display the model summary. Make a scatter plot of `temp` versus `year`, and add the fitted curves from the linear and 3rd order polynomial models to the graph. One way to do this is with `lines` command after creating the plot, e.g. `lines(mod$fitted~year, data=aatemp)`. Which model seems to track the data better, based on what you see?

Answer:

```
armodel3 = gls(temp ~ year + I(year^2) + I(year^3),
               correlation = corAR1(form = ~ year), data = aatemp)
summary(armodel3)
```

```
## Generalized least squares fit by REML
##   Model: temp ~ year + I(year^2) + I(year^3)
##   Data: aatemp
##        AIC      BIC   logLik
##   466.35 482.6071 -227.175
##
## Correlation Structure: ARMA(1,0)
##  Formula: ~year
##  Parameter estimate(s):
##      Phi1
## 0.214179
##
## Coefficients:
##                Value Std.Error   t-value p-value
## (Intercept) 41531.81 20557.977  2.020228  0.0458
## year          -64.59    31.956 -2.021152  0.0457
## I(year^2)       0.03     0.017  2.023909  0.0454
## I(year^3)       0.00     0.000 -2.026180  0.0451
##
##  Correlation:
##            (Intr) year I(y^2)
## year       -1
## I(year^2)   1     -1
## I(year^3) -1       1    -1
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.6468228 -0.6658191 -0.1092529  0.7129614  2.8531996
##
## Residual standard error: 1.456044
## Degrees of freedom: 115 total; 111 residual
```

```
plot(aatemp$year, aatemp$temp)
lines(armodel$fitted ~ year, data = aatemp, col = 4)
lines(armodel3$fitted ~ year, data = aatemp, col = 6)
```



As we can see, the cubic one tracks the model much better.

Problem 3:

We delve into the theory for added variable plots and variance inflation factors. Consider a model of the form

$$\mathbf{y} = \mathbf{X}_0\mathbf{fi} + \mathbf{z}\gamma + \mathbf{e}$$

where $\mathbf{X_0}$ is $n \times p$ and full rank, $\mathbf{z}$ is $n \times 1$ and linearly independent of the columns of $\mathbf{X_0}$, $E(\mathbf{e}) = \mathbf{0}$, and $Cov(\mathbf{e}) = \sigma^2\mathbf{I}$.

The partial regression plot for $\mathbf{z}$ shows its association with the response $\mathbf{y}$ after adjusting for the variables in $\mathbf{X_0}$. This plot is obtained by plotting residuals $\mathbf{r}_y$ from the LS regression of $\mathbf{y}$ on $\mathbf{X}_0$ versus residuals $\mathbf{r}_z$ from the LS regression of $\mathbf{z}$ on $\mathbf{X}_0$.

a) Show that $\mathbf{r}_y = (\mathbf{I} - \mathbf{H}_0)\mathbf{y}$ and $\mathbf{r_z} = (\mathbf{I} - \mathbf{H}_0)\mathbf{z}$, where $\mathbf{H}_0 = \mathbf{X_0}(\mathbf{X_0^T X_0})^{-\mathbf{1}}\mathbf{X_0^T}$.

Answer: Recall $\hat{\mathbf{y}} = \mathbf{X}_0\hat{\boldsymbol{\beta}} = \mathbf{X}_0(\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T\mathbf{y}$. Therefore,

$$\begin{aligned}\mathbf{r}_y &= \mathbf{y} - \mathbf{X}_0(\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T\mathbf{y} \\ &= (\mathbf{I} - \mathbf{X}_0(\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T)\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H}_0)\mathbf{y}.\end{aligned}$$

Similarly,

$$\mathbf{r}_z = \mathbf{z} - \mathbf{X}_0(\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T\mathbf{z} = (\mathbf{I} - \mathbf{H}_0)\mathbf{z},$$

which finishes the proof.

b) Under the model assumptions given above, show that $E(\mathbf{r}_y) = \mathbf{r}_z\gamma$, so the expected slope of the line is the same as the coefficient of $\mathbf{z}$ in the full model.

Answer: We compute from a)

$$\begin{aligned}
E(\mathbf{r}_y) &= E\big((\mathbf{I}-\mathbf{H}_0)\mathbf{y}\big) \\
&= (\mathbf{I}-\mathbf{H}_0)E(\mathbf{y}) && (\mathbf{I}-\mathbf{H}_0 \text{ is constant}) \\
&= (\mathbf{I}-\mathbf{H}_0)(\mathbf{X}_0\boldsymbol{\beta}+\mathbf{z}\gamma+E(\mathbf{e})) \\
&= (\mathbf{I}-\mathbf{H}_0)(\mathbf{X}_0\boldsymbol{\beta}+\mathbf{z}\gamma) && (E(\mathbf{e})=\mathbf{0}) \\
&= (\mathbf{I}-\mathbf{H}_0)\mathbf{z}\gamma && (\mathbf{X}_0\boldsymbol{\beta}\in \mathrm{col}(\mathbf{X}_0)) \\
&= \mathbf{r}_z\gamma, && (\text{from a}))
\end{aligned}$$

where $\mathrm{col}(\mathbf{X}_0)$ denotes the column space of $\mathbf{X}_0$ and we use the fact that projection matrix of the space, to which a vector is orthogonal, i.e., $\mathbf{I}-\mathbf{H}_0$, projects this vector into a zero vector.

c) The conditional expectation model in b) has the form of simple linear regression through the origin (no intercept). Show that fitting this "model" by LS regression gives estimated slope:

$$\hat{\gamma} = \frac{\mathbf{z}^{\mathbf{T}}(\mathbf{I}-\mathbf{H}_{\mathbf{0}})\mathbf{y}}{\mathbf{z}^{\mathbf{T}}(\mathbf{I}-\mathbf{H}_{\mathbf{0}})\mathbf{z}} = \frac{\sum_{i=1}^{n}(z_i-\hat{z}_i)y_i}{\sum_{i=1}^{n}(z_i-\hat{z}_i)^2}$$

where $\hat{\mathbf{z}} = \mathbf{H}_0\mathbf{z}$ and y_i, z_i and $\hat{z}_i$ are the ith components of $\mathbf{y}$, $\mathbf{z}$ and $\hat{\mathbf{z}}$ respectively.

Answer: Recall the formula of OLS estimates for the model of regression from origin from a) in Problem 2 in HW1. Here the covariates is $\mathbf{r}_z$ and the response is $\mathbf{r}_y$. Thus, we can plug in those quantities into the formula to obtain

$$\begin{aligned}
\hat{\gamma} = \frac{\mathbf{r}_z^T\mathbf{r}_y}{\mathbf{r}_z^T\mathbf{r}_z} &= \frac{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)^T(\mathbf{I}-\mathbf{H}_0)\mathbf{y}}{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z}} \\
&= \frac{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{y}}{\hat{\mathbf{z}}^T(\mathbf{I}-\mathbf{H}_0)^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z}} \\
&= \frac{\sum_{i=1}^{n}(z_i-\hat{z}_i)y_i}{\sum_{i=1}^{n}(z_i-\hat{z}_i)^2},
\end{aligned}$$

where we applied the fact that for a projection matrix $\mathbf{P}$, $\mathbf{P}=\mathbf{P}^T$ and $\mathbf{PP}=\mathbf{P}$.

d) Using c) and the model assumptions, show $E(\hat{\gamma}) = \gamma$ and

$$var(\hat{\gamma}) = \frac{\sigma^2}{\sum_{i=1}^{n}(z_i-\hat{z}_i)^2}$$

Answer: Since $\mathbf{z}$ is deterministic, from c), we compute

$$\begin{aligned} E(\hat{\gamma}) = \frac{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)E(\mathbf{y})}{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z}} &= \frac{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)(\mathbf{X}_0\boldsymbol{\beta}+\mathbf{z}\boldsymbol{\gamma})}{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z}} \\ &= \frac{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z}}{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z}} \cdot \gamma \\ &= \gamma, \end{aligned}$$

and

$$\begin{aligned} var(\hat{\gamma}) = \frac{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)var(\mathbf{y})(\mathbf{I}-\mathbf{H}_0)\mathbf{z}}{(\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z})^2} &= \frac{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{I}(\mathbf{I}-\mathbf{H}_0)\mathbf{z}}{(\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z})^2} \cdot \sigma^2 \\ &= \frac{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z}}{(\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z})^2} \cdot \sigma^2 \\ &= \frac{\sigma^2}{\mathbf{z}^T(\mathbf{I}-\mathbf{H}_0)\mathbf{z}} \\ &= \frac{\sigma^2}{\sum_{i=1}^n (z_i - \hat{z}_i)^2}, \end{aligned}$$

which finishes the proof.

e) Let R_z^2 denote the multiple R-square statistic for the regression of $\mathbf{z}$ on the variables in $\mathbf{X}_0$. It can be shown that

$$R_z^2 = 1 - \frac{\sum_{i=1}^n (z_i - \hat{z}_i)^2}{\sum_{i=1}^n (z_i - \bar{z}_i)^2}$$

where $\bar{z} = n^{-1}\sum_{i=1}^n z_i$. Using this fact, show that

$$var(\hat{\gamma}) = VIF_z * \frac{\sigma^2}{\sum_{i=1}^n (z_i - \bar{z}_i)^2}$$

where VIF_z is the "variance inflation factor" given by

$$VIF_z = \frac{1}{1-R_z^2}$$

Answer: From d), we compute,

$$\begin{aligned} var(\hat{\gamma}) &= \frac{\sigma^2}{\sum_{i=1}^n (z_i - \hat{z}_i)^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^n (z_i - \bar{z})^2} \cdot \frac{\sum_{i=1}^n (z_i - \bar{z})^2}{\sum_{i=1}^n (z_i - \hat{z}_i)^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^n (z_i - \bar{z})^2} \cdot \frac{1}{1-R_z^2} \\ &= VIF_z \times \frac{\sigma^2}{\sum_{i=1}^n (z_i - \bar{z})^2}, \end{aligned}$$

which finishes the proof.

STAT 425 Assignment 5

Due Tuesday, April 6, 11:59 pm. Submit through Moodle.

Name: (insert your name here)

Netid: (insert)

Submit your computational work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown.

Most relevant class notes: 5.2.Spline, 6.Ancova, 7.VarSelect

Problem 1

Consider the `prostate` cancer surgery data from the `faraway` library in **R**. The variable `lpsa` is a measurement of prostate specific antigen on the log scale. Treat `lpsa` as the response and all the other variables in the data frame as potential predictors.

a) Using backward elimination testing and an alpha cutoff of 0.10, find the best model according to this procedure. Be sure to include the steps and the final model.

Answer:

```
library(faraway)

fit_full = lm(lpsa ~ ., data = prostate)
summary(fit_full)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

```r
#The gleason coefficient has the highest p-value above alpha = 0.10.

fit2 = lm(lpsa ~ . -gleason, data = prostate)
summary(fit2)
```

```
##
## Call:
## lm(formula = lpsa ~ . - gleason, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.73117 -0.38137 -0.01728  0.43364  1.63513
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.953926   0.829439   1.150  0.25319
## lcavol       0.591615   0.086001   6.879 8.07e-10 ***
## lweight      0.448292   0.167771   2.672  0.00897 **
## age         -0.019336   0.011066  -1.747  0.08402 .
## lbph         0.107671   0.058108   1.853  0.06720 .
## svi          0.757734   0.241282   3.140  0.00229 **
## lcp         -0.104482   0.090478  -1.155  0.25127
## pgg45        0.005318   0.003433   1.549  0.12488
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7048 on 89 degrees of freedom
## Multiple R-squared:  0.6544, Adjusted R-squared:  0.6273
## F-statistic: 24.08 on 7 and 89 DF,  p-value: < 2.2e-16
```

```r
#The lcp coefficient has the highest p-value above alpha = 0.10.

fit3 = lm(lpsa ~ . - gleason - lcp, data = prostate)
summary(fit3)
```

```
##
## Call:
## lm(formula = lpsa ~ . - gleason - lcp, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77711 -0.41708  0.00002  0.40676  1.59681
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.980085   0.830665   1.180  0.24116
## lcavol       0.545770   0.076431   7.141 2.31e-10 ***
## lweight      0.449450   0.168078   2.674  0.00890 **
## age         -0.017470   0.010967  -1.593  0.11469
## lbph         0.105755   0.058191   1.817  0.07249 .
## svi          0.641666   0.219757   2.920  0.00442 **
## pgg45        0.003528   0.003068   1.150  0.25331
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7061 on 90 degrees of freedom
## Multiple R-squared:  0.6493, Adjusted R-squared:  0.6259
## F-statistic: 27.77 on 6 and 90 DF,  p-value: < 2.2e-16
```

```r
#The pgg45 coefficient has the highest p-value above alpha = 0.10.

fit4 = lm(lpsa ~ . - gleason - lcp - pgg45, data = prostate)
summary(fit4)
```

```
##
## Call:
## lm(formula = lpsa ~ . - gleason - lcp - pgg45, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.83505 -0.39396  0.00414  0.46336  1.57888
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.95100    0.83175   1.143 0.255882
## lcavol       0.56561    0.07459   7.583 2.77e-11 ***
## lweight      0.42369    0.16687   2.539 0.012814 *
## age         -0.01489    0.01075  -1.385 0.169528
## lbph         0.11184    0.05805   1.927 0.057160 .
## svi          0.72095    0.20902   3.449 0.000854 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7073 on 91 degrees of freedom
## Multiple R-squared:  0.6441, Adjusted R-squared:  0.6245
## F-statistic: 32.94 on 5 and 91 DF,  p-value: < 2.2e-16
```

```
#The age coefficient has the highest p-value above alpha = 0.10.

fit5 = lm(lpsa ~ . - gleason - lcp - pgg45 - age, data = prostate)
summary(fit5)
```

```
##
## Call:
## lm(formula = lpsa ~ . - gleason - lcp - pgg45 - age, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82653 -0.42270  0.04362  0.47041  1.48530
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.14554    0.59747   0.244  0.80809
## lcavol       0.54960    0.07406   7.422 5.64e-11 ***
## lweight      0.39088    0.16600   2.355  0.02067 *
## lbph         0.09009    0.05617   1.604  0.11213
## svi          0.71174    0.20996   3.390  0.00103 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7108 on 92 degrees of freedom
## Multiple R-squared:  0.6366, Adjusted R-squared:  0.6208
## F-statistic: 40.29 on 4 and 92 DF,  p-value: < 2.2e-16
```

```
#The lbph coefficient has the highest p-value above alpha = 0.10.

fit6 = lm(lpsa ~ . - gleason - lcp - pgg45 - age - lbph, data = prostate)
summary(fit6)
```

```
##
## Call:
```

```
## lm(formula = lpsa ~ . - gleason - lcp - pgg45 - age - lbph, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.72964 -0.45764  0.02812  0.46403  1.57013
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.26809    0.54350  -0.493  0.62298
## lcavol       0.55164    0.07467   7.388  6.3e-11 ***
## lweight      0.50854    0.15017   3.386  0.00104 **
## svi          0.66616    0.20978   3.176  0.00203 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7168 on 93 degrees of freedom
## Multiple R-squared:  0.6264, Adjusted R-squared:  0.6144
## F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

```
#All of these predictors are significant at the alpha = .10 level, so this is the fina
```

b) Use backward selection in a stepwise algorithm to find the best model according to the AIC criterion. Be sure to include the steps and the final model.

Answer:

```
step(fit_full, direction = "backward")
```

```
## Start:  AIC=-58.32
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##
##           Df Sum of Sq    RSS     AIC
## - gleason  1    0.0412 44.204 -60.231
## - pgg45    1    0.5258 44.689 -59.174
## - lcp      1    0.6740 44.837 -58.853
## <none>                 44.163 -58.322
## - age      1    1.5503 45.713 -56.975
## - lbph     1    1.6835 45.847 -56.693
## - lweight  1    3.5861 47.749 -52.749
## - svi      1    4.9355 49.099 -50.046
## - lcavol   1   22.3721 66.535 -20.567
##
## Step:  AIC=-60.23
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - lcp      1    0.6623 44.867 -60.789
## <none>                 44.204 -60.231
## - pgg45    1    1.1920 45.396 -59.650
## - age      1    1.5166 45.721 -58.959
## - lbph     1    1.7053 45.910 -58.560
## - lweight  1    3.5462 47.750 -54.746
## - svi      1    4.8984 49.103 -52.037
## - lcavol   1   23.5039 67.708 -20.872
##
## Step:  AIC=-60.79
## lpsa ~ lcavol + lweight + age + lbph + svi + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - pgg45    1    0.6590 45.526 -61.374
## <none>                 44.867 -60.789
## - age      1    1.2649 46.131 -60.092
## - lbph     1    1.6465 46.513 -59.293
## - lweight  1    3.5647 48.431 -55.373
## - svi      1    4.2503 49.117 -54.009
## - lcavol   1   25.4189 70.285 -19.248
##
## Step:  AIC=-61.37
## lpsa ~ lcavol + lweight + age + lbph + svi
##
##           Df Sum of Sq    RSS     AIC
## <none>                 45.526 -61.374
## - age      1    0.9592 46.485 -61.352
## - lbph     1    1.8568 47.382 -59.497
## - lweight  1    3.2251 48.751 -56.735
## - svi      1    5.9517 51.477 -51.456
## - lcavol   1   28.7665 74.292 -15.871
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi, data = prostate)
##
## Coefficients:
## (Intercept)       lcavol      lweight          age         lbph          svi
##     0.95100      0.56561      0.42369     -0.01489      0.11184      0.72095
```

Step 1: gleason is removed. Step 2: lcp is removed. Step 3: pgg45 is removed. Full model has lcavol, lweight, age, lbph, and svi as predictors.

c) Use stepwise selection with the "both" option to find the best model according to the BIC criterion. Include the steps and the final model.

Answer:

```r
n = length(prostate[,1])
step(fit_full, direction = "both", k = log(n))
```

```
## Start:  AIC=-35.15
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##
##           Df Sum of Sq    RSS     AIC
## - gleason  1    0.0412 44.204 -39.634
## - pgg45    1    0.5258 44.689 -38.576
## - lcp      1    0.6740 44.837 -38.255
## - age      1    1.5503 45.713 -36.377
## - lbph     1    1.6835 45.847 -36.095
## <none>                 44.163 -35.149
## - lweight  1    3.5861 47.749 -32.151
## - svi      1    4.9355 49.099 -29.448
## - lcavol   1   22.3721 66.535   0.030
##
## Step:  AIC=-39.63
## lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - lcp      1    0.6623 44.867 -42.766
## - pgg45    1    1.1920 45.396 -41.627
## - age      1    1.5166 45.721 -40.936
## - lbph     1    1.7053 45.910 -40.537
## <none>                 44.204 -39.634
## - lweight  1    3.5462 47.750 -36.723
## + gleason  1    0.0412 44.163 -35.149
## - svi      1    4.8984 49.103 -34.014
## - lcavol   1   23.5039 67.708  -2.849
##
## Step:  AIC=-42.77
## lpsa ~ lcavol + lweight + age + lbph + svi + pgg45
##
##           Df Sum of Sq    RSS     AIC
## - pgg45    1    0.6590 45.526 -45.926
## - age      1    1.2649 46.131 -44.644
## - lbph     1    1.6465 46.513 -43.844
## <none>                 44.867 -42.766
## - lweight  1    3.5647 48.431 -39.925
## + lcp      1    0.6623 44.204 -39.634
## - svi      1    4.2503 49.117 -38.561
## + gleason  1    0.0296 44.837 -38.255
## - lcavol   1   25.4189 70.285  -3.800
##
## Step:  AIC=-45.93
## lpsa ~ lcavol + lweight + age + lbph + svi
##
##           Df Sum of Sq    RSS     AIC
## - age      1    0.9592 46.485 -48.478
## - lbph     1    1.8568 47.382 -46.623
## <none>                 45.526 -45.926
## - lweight  1    3.2251 48.751 -43.862
## + pgg45    1    0.6590 44.867 -42.766
## + gleason  1    0.4560 45.070 -42.328
## + lcp      1    0.1293 45.396 -41.627
## - svi      1    5.9517 51.477 -38.583
## - lcavol   1   28.7665 74.292  -2.997
##
## Step:  AIC=-48.48
## lpsa ~ lcavol + lweight + lbph + svi
##
##           Df Sum of Sq    RSS     AIC
## - lbph     1    1.3001 47.785 -50.377
## <none>                 46.485 -48.478
## - lweight  1    2.8014 49.286 -47.377
## + age      1    0.9592 45.526 -45.926
## + pgg45    1    0.3533 46.131 -44.644
## + gleason  1    0.2126 46.272 -44.348
## + lcp      1    0.1023 46.383 -44.117
## - svi      1    5.8063 52.291 -41.636
## - lcavol   1   27.8298 74.315  -7.542
##
## Step:  AIC=-50.38
## lpsa ~ lcavol + lweight + svi
##
##           Df Sum of Sq    RSS     AIC
## <none>                 47.785 -50.377
## + lbph     1    1.3001 46.485 -48.478
## + pgg45    1    0.5735 47.211 -46.974
## + age      1    0.4025 47.382 -46.623
## + gleason  1    0.3890 47.396 -46.596
## + lcp      1    0.0641 47.721 -45.933
## - svi      1    5.1814 52.966 -44.966
## - lweight  1    5.8924 53.677 -43.673
## - lcavol   1   28.0445 75.829 -10.160

##
```

```
## Call:
## lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)
##
## Coefficients:
## (Intercept)       lcavol      lweight          svi
##     -0.2681       0.5516       0.5085       0.6662
```

Step 1: gleason is removed. Step 2: lcp is removed. Step 3: pgg45 is removed. Step 4: age is removed. Step 5: lbph is removed. Full model includes lcavol, lweight, and svi as predictors.

d) Use the leaps and bounds algorithm to determine the model with smallest residual sums of squares for each model size from 2 to the maximum possible based on the number of columns in the data frame. Display the results by showing "which" variables were selected for model sizes 2, 3, etc.

Answer:

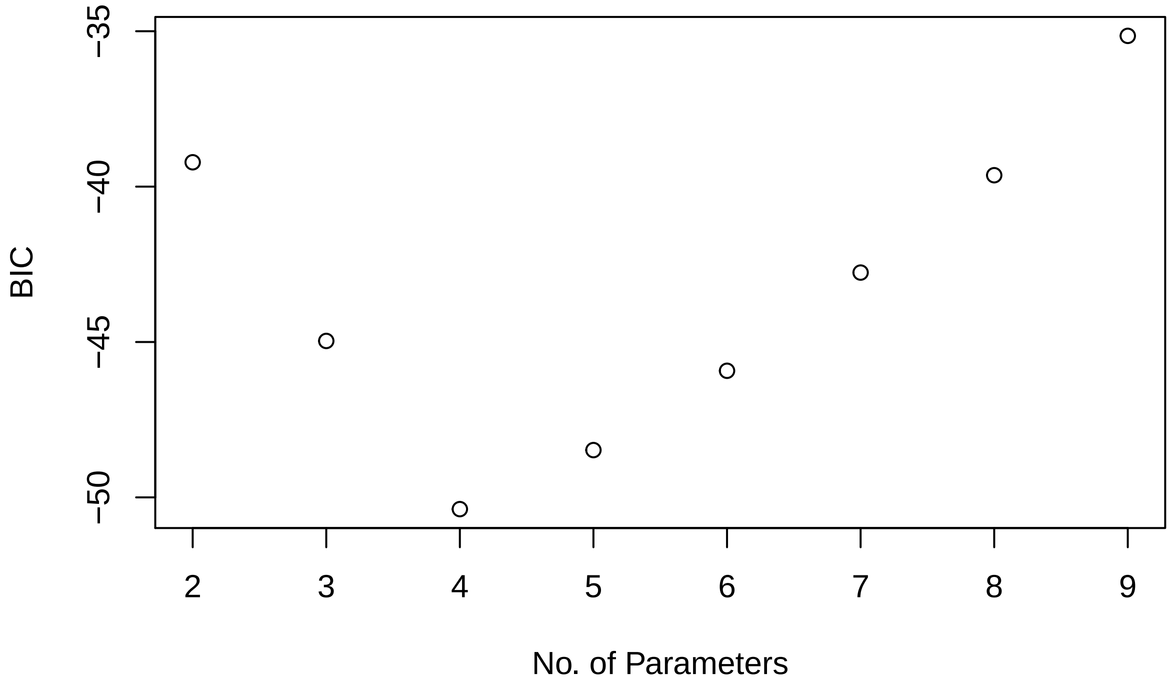
```
library(leaps)
b=regsubsets(lpsa ~ ., data = prostate)
rs = summary(b)
rs$which
```

```
##   (Intercept) lcavol lweight   age  lbph   svi   lcp gleason pgg45
## 1        TRUE   TRUE   FALSE FALSE FALSE FALSE FALSE   FALSE FALSE
## 2        TRUE   TRUE    TRUE FALSE FALSE FALSE FALSE   FALSE FALSE
## 3        TRUE   TRUE    TRUE FALSE FALSE  TRUE FALSE   FALSE FALSE
## 4        TRUE   TRUE    TRUE FALSE  TRUE  TRUE FALSE   FALSE FALSE
## 5        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE FALSE   FALSE FALSE
## 6        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE FALSE   FALSE  TRUE
## 7        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE  TRUE   FALSE  TRUE
## 8        TRUE   TRUE    TRUE  TRUE  TRUE  TRUE  TRUE    TRUE  TRUE
```

e) Using the results from d) and further calculations, graph BIC versus model size for the models selected in part d). Which model is the overall best model according to BIC?

Answer:

```
msize = 2:9
Bic = n*log(rs$rss/n) + msize*log(n)
plot(msize, Bic, xlab="No. of Parameters", ylab = "BIC")
```



As shown by the graph, model size 4, with an intercept and three predictors has the lowest BIC. In the table shown in part d, a model with lcavol, lweight, and svi as predictors is the best overall model according to BIC.

Answer:

Problem 2

The `aatemp` data in the `faraway` library comes from the U.S. Historical Climatological Network. The data report annual mean temperatures in Ann Arbor Michigan for roughly 150 years.

a) With `temp` as the response, fit a regression spline with intercept using B-spline basis functions of `year` and 8 degrees of freedom. Show the fitted curve on the scatter plot of `temp` versus `year`.

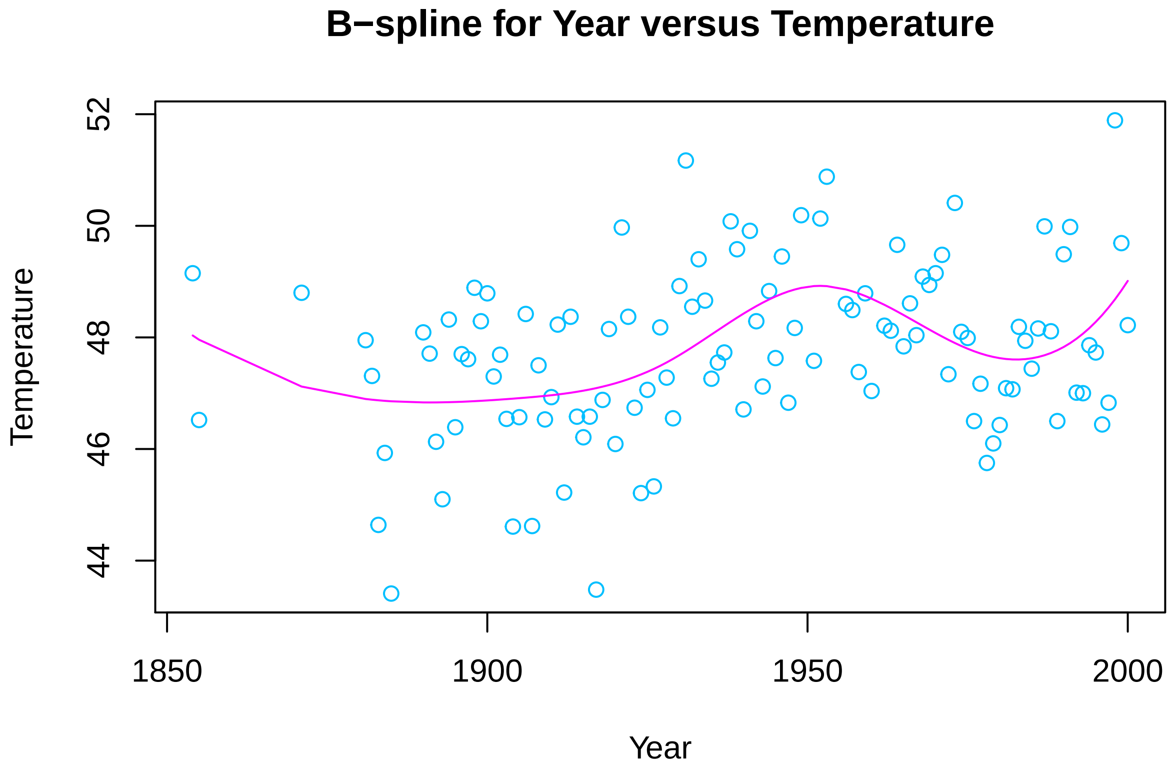
Answer: We first create a basis function for `year` using `bs` and plot the fitted curve as below.

```
# load package and data
require('faraway')
require('splines')
```

```
## Loading required package: splines
```

```
data('aatemp')
# B-spline
bs.basis <- bs(aatemp$year, df = 8, intercept = TRUE)
# regression
```

```r
lmod <- lm(aatemp$temp ~ bs.basis - 1)
# plot
plot(temp ~ year, data = aatemp, col = 'deepskyblue',
     xlab = 'Year', ylab = 'Temperature',
     main = 'B-spline for Year versus Temperature')
lines(x = aatemp$year, y = fitted.values(lmod), col = 'magenta')
```



b) How many knots does the model in a) have?

Answer: Based on the formula df $= m + 4$ for B-spline with an intercept, where m is the number of knots. We know that we have $m = 4$ knots for model in a).

c) Compute AIC, BIC and adjusted R-square for the model in a).

Answer: We can compute AIC, BIC, and adjusted R-square as below.

```r
# AIC
AIC(lmod)
```

```
## [1] 414.0364
```

```r
# BIC
BIC(lmod)
```

```
## [1] 438.7408
```

```r
# Adjusted R square
n <- nrow(aatemp)
```

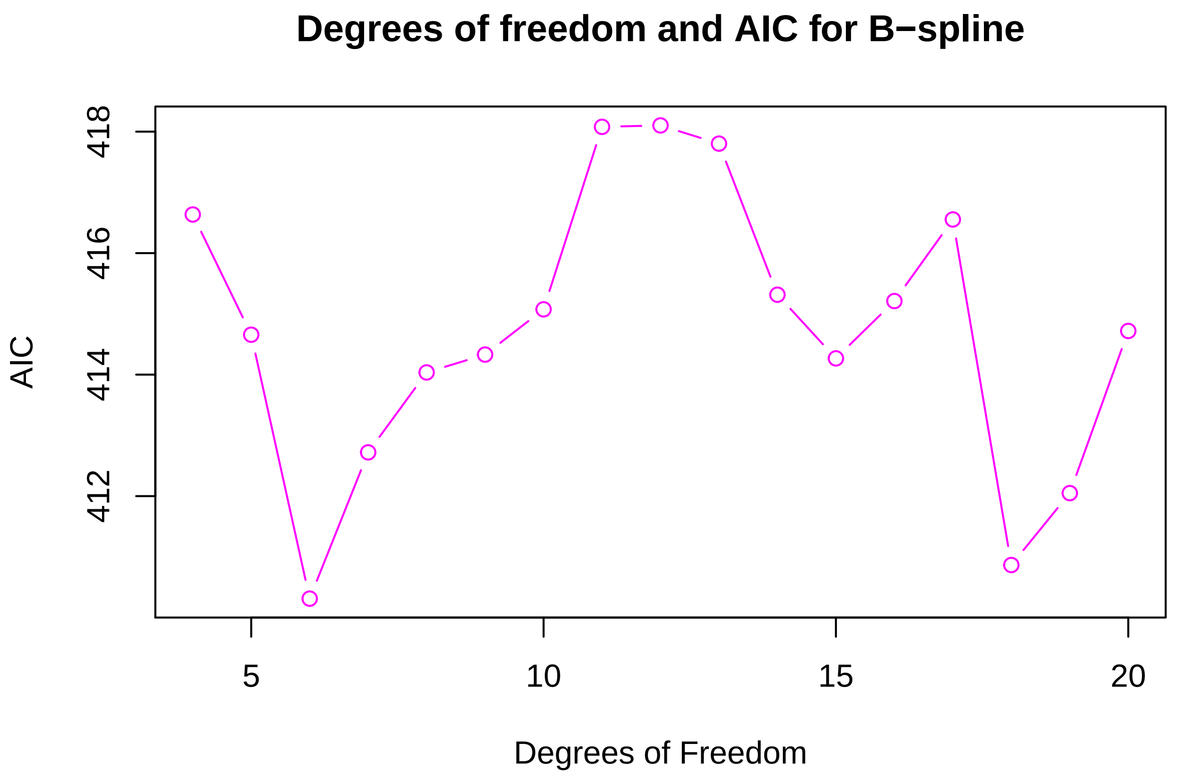
```
sse <- as.numeric(crossprod(residuals(lmod)))
sst <- var(aatemp$temp) * (nrow(aatemp) - 1)
r2 <- 1 - sse / sst
adj.r2 <- 1 - (1 - r2) * (n - 1) / (n - 8)
```

Note that if you calculate the AIC, BIC, and adjusted R^2 by yourself, your answer should be correct if the formula are applied correctly. Also, if you use p or $p + 1$ in the computation of AIC, where p is the dimension of the design matrix, you should also be fine.

d) Compute AIC for the b-spline models with degrees of freedom 4, 5, 6, . . . 20. Plot AIC versus degrees of freedom. Which of these models is the best, according to AIC?

Answer: We can write a loop as below to compute the AIC and find which model is best according to AIC.

```
# loop
DF <- 4:20
AICs <- c()
for (df in DF) {
  # B-spline
  bs.basis <- bs(aatemp$year, df = df, intercept = TRUE)
  # regression
  lmod <- lm(aatemp$temp ~ bs.basis)
  # AIC
  AICs <- c(AICs, AIC(lmod))
}
# plot
plot(AICs ~ DF, xlab = 'Degrees of Freedom',
     ylab = 'AIC',
     main = 'Degrees of freedom and AIC for B-spline', col = 'magenta',
     type = 'b')
```



```
# which minimizes
DF[which.min(AICs)]
```

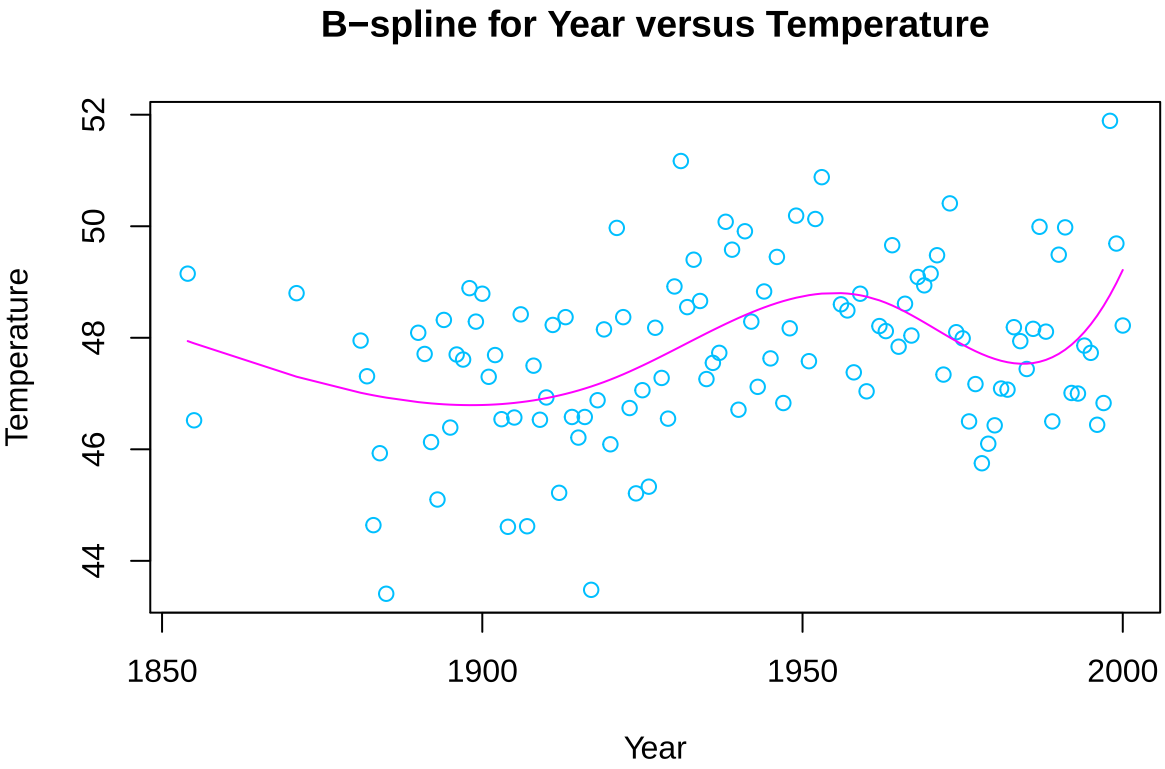
```
## [1] 6
```

As we can see, the B-spline model is degrees of freedom of 6 is considered the best according to AIC.

e) Show the fitted curve from the best model selected in d) on the scatter plot of `temp` versus `year`. How does it compare with the curve in a)?

Answer: We can plot the fitted curve of the best model in d) as below.

```
# B-spline
bs.basis <- bs(aatemp$year, df = 6, intercept = TRUE)
# regression
lmod <- lm.fit(x = bs.basis, y = aatemp$temp)
# plot
plot(temp ~ year, data = aatemp, col = 'deepskyblue',
     xlab = 'Year', ylab = 'Temperature',
     main = 'B-spline for Year versus Temperature')
lines(x = aatemp$year, y = fitted.values(lmod), col = 'magenta')
```



As we can see, compared to a), the plot is almost identical.

Problem 3:

In this problem we model data from a study of the prevalence of obesity, diabetes and cardiovascular disease among 403 African Americans in central Virginia. The data are in the `diabetes` dataset in the `faraway` library. One of the blood measurements is glycosolated hemoglobin (`glyhb`). A value higher than 7 is often considered to be a positive diagnosis for diabetes. Here we treat the numerical value og `glyhb` as the response, and consider possible predictor variables `gender` (a factor variable), `waist`, `age`, and stabilized glucose (`stab.glu`).

a) Fit the ancova model that includes the main effects for `gender`, `waist`, `age`, `stab.glu` as well as the two-way interactions between gender and each of the other three predictor variables. Show the model summary and indicate which coefficients are significant at level 0.05.

Answer:

```
library("faraway")
model1 = lm(glyhb ~ gender + waist + age + stab.glu + gender:waist + gender:age + gender
summary(model1)
```

```
##
## Call:
## lm(formula = glyhb ~ gender + waist + age + stab.glu + gender:waist +
##     gender:age + gender:stab.glu, data = diabetes)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2562 -0.7162 -0.1203  0.4528  9.6957
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.393920   0.838982   2.853  0.00456 **
## genderfemale         -2.492367   1.058500  -2.355  0.01905 *
## waist                -0.011856   0.022109  -0.536  0.59210
## age                   0.013983   0.007507   1.863  0.06328 .
## stab.glu              0.027386   0.001891  14.484  < 2e-16 ***
## genderfemale:waist    0.048418   0.027563   1.757  0.07979 .
## genderfemale:age      0.003624   0.009664   0.375  0.70789
## genderfemale:stab.glu 0.005616   0.003001   1.872  0.06203 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.45 on 380 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.5898, Adjusted R-squared:  0.5822
## F-statistic: 78.06 on 7 and 380 DF,  p-value: < 2.2e-16
```

Based on the p values of summary result, we can see coefficients of "gender" and "stab.glu" are significant.

b) By comparing with a simplified model, test the null hypothesis that the three interaction coefficients all equal zero versus the alternative that at least one of them is nonzero. Use overall significance level of 0.05 and state your conclusion.

Answer:

```
model2 = lm(glyhb ~ gender + waist + age + stab.glu, data = diabetes)
anova(model1, model2)
```

```
## Analysis of Variance Table
##
## Model 1: glyhb ~ gender + waist + age + stab.glu + gender:waist + gender:age +
##     gender:stab.glu
## Model 2: glyhb ~ gender + waist + age + stab.glu
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1    380 798.52
## 2    383 818.95 -3   -20.432 3.241 0.02216 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the p value we should reject the null hypothesis, which means at least one of interaction coefficients is nonzero.

c) The additive model has only the main effects; no interactions are included. (Main effects are terms involving only the original variables, not their products). Fit this model and determine which of the main effects appear to have statistically significant coefficients at the 0.05 level.

Answer:

```
modeladd = lm(glyhb ~ gender + waist + age + stab.glu, data = diabetes)
summary(modeladd)
```

```
##
## Call:
## lm(formula = glyhb ~ gender + waist + age + stab.glu, data = diabetes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7806 -0.7293 -0.1601  0.4072  9.6192
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.835159   0.520004   1.606 0.109084
## genderfemale 0.109011   0.151993   0.717 0.473683
## waist        0.020016   0.013315   1.503 0.133584
## age          0.017042   0.004755   3.584 0.000382 ***
## stab.glu     0.029261   0.001474  19.851  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.462 on 383 degrees of freedom
##   (15 observations deleted due to missingness)
## Multiple R-squared:  0.5793, Adjusted R-squared:  0.5749
## F-statistic: 131.9 on 4 and 383 DF,  p-value: < 2.2e-16
```

From the result we can see the variables “age” and “stab.glu” are two significant coefficients.

d) Compute AIC for the model in a) and the model in c). Which model is preferred according to this criterion?

Answer:

```
AIC(model1)
```

```
## [1] 1399.137
```

```
AIC(modeladd)
```

```
## [1] 1402.94
```

As we can see, the model in a) is prefered here since the AIC value is smaller.

e) Use the sequential analysis of variance table for the full model in a) to determine which main effects and interactions are significant. What do you conclude?

Answer:

```
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: glyhb
##                  Df Sum Sq Mean Sq  F value    Pr(>F)    
## gender            1   4.81    4.81   2.2875   0.13125    
## waist             1 101.82  101.82  48.4538 1.492e-11 ***
## age               1 178.50  178.50  84.9424 < 2.2e-16 ***
## stab.glu          1 842.61  842.61 400.9834 < 2.2e-16 ***
## gender:waist      1  11.37   11.37   5.4109   0.02054 *  
## gender:age        1   1.70    1.70   0.8094   0.36887    
## gender:stab.glu   1   7.36    7.36   3.5028   0.06203 .  
## Residuals       380 798.52    2.10                       
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the sequential anova test, we can see that all of three main effects and interaction between "gender" and "waist" are significant.

STAT 425 Assignment 6

Due Monday, April 19, 11:59 pm. Submit through Moodle.

Name: (insert your name here)

Netid: (insert)

Submit your computational work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown.

Most relevant class notes: 8.Shrinkage, R_Shrink.Rmd, 9.1.OneWayAnova1, 9.2.OneWayAnova2. We also use some of our methods from earlier in the class.

Problem 1

Consider the `fat` data from the `faraway` library in **R**. The following code is an example of how to select a random test set of 25 observations, and to use the remaining observations as the training set. In the code, we set the random seed to make the result reproducible, but this seed can be changed.

```
library(faraway)
n=dim(fat)[1]
set.seed(12357)
testid = sample(n, 25, replace=FALSE)
trainid = -testid
test = fat[testid,]
train = fat[trainid,]
```

We will compare several regression methods using train/test evaluation.

a) For the `fat` data, create a randomly selected test set of 25 observations and a training set consisting of all the other observations, removing the variables `brozek` and `density` from the data. Display the first 6 rows of the training and test sets. Also display the dimensions of the training data frame and test data frame.

Answer:

```
fattrain = train[ ,-c(1, 3)]
fattest = test[ ,-c(1,3)]
```

First six rows of the “train”:

```
fattrain[1:6, ]
```

```
##   siri age weight height adipos  free neck chest abdom   hip thigh knee ankle
## 1 12.3  23 154.25  67.75   23.7 134.9 36.2  93.1  85.2  94.5  59.0 37.3  21.9
## 2  6.1  22 173.25  72.25   23.4 161.3 38.5  93.6  83.0  98.7  58.7 37.3  23.4
## 3 25.3  22 154.00  66.25   24.7 116.0 34.0  95.8  87.9  99.2  59.6 38.9  24.0
## 4 10.4  26 184.75  72.25   24.9 164.7 37.4 101.8  86.4 101.2  60.1 37.3  22.8
## 5 28.7  24 184.25  71.25   25.6 133.1 34.4  97.3 100.0 101.9  63.2 42.2  24.0
## 6 20.9  24 210.25  74.75   26.5 167.0 39.0 104.5  94.4 107.8  66.0 42.0  25.6
##   biceps forearm wrist
## 1   32.0    27.4  17.1
## 2   30.5    28.9  18.2
## 3   28.8    25.2  16.6
## 4   32.4    29.4  18.2
## 5   32.2    27.7  17.7
## 6   35.7    30.6  18.8
```

Dimensions of “train”:

```
dim(fattrain)
```

```
## [1] 227  16
```

First six rows of the “test”:

```
fattest[1:6, ]
```

```
##     siri age weight height adipos  free neck chest abdom   hip thigh knee ankle
## 91  20.5  46 177.00  70.00   25.4 141.3 37.2  99.7  95.6 102.2  58.3 38.2  22.5
## 95   9.0  47 184.25  74.50   23.4 166.6 37.3  99.6  88.8 101.4  57.4 39.6  24.6
## 36  40.1  49 191.75  65.00   32.0 118.4 38.4 118.5 113.1 113.8  61.9 38.3  21.9
## 43  31.6  48 217.00  70.00   31.2 151.1 37.3 113.3 111.2 114.1  67.7 40.9  25.0
## 175 25.3  36 226.75  71.75   31.0 170.9 41.5 115.3 108.8 114.4  69.2 42.4  24.0
## 121 27.9  52 206.50  74.50   26.2 150.7 40.8 104.3  99.2 104.1  58.5 39.3  24.6
##     biceps forearm wrist
## 91    29.1    27.7  17.7
## 95    30.3    27.9  17.8
## 36    32.0    29.8  17.0
## 43    36.7    29.8  18.4
## 175   35.4    21.0  20.1
## 121   33.9    31.2  19.5
```

Dimensions of “test”:

```
dim(fattest)
```

```
## [1] 25 16
```

b) Use the training data to estimate the linear regression of `siri` on all of the other variables except for `brozek` and `density`. Then use the test data to compute the estimated mean square error for prediction.

Answer:

```
mse <- function(y1, y2)
{
  mean((y1 - y2)^2)
}
```

```
lmmodel = lm(siri ~ ., data = fattrain)
mse( predict(lmmodel, newdata = fattest), fattest$siri)
```

```
## [1] 2.090238
```

c) Repeat exercise b) for linear regression with variables selected using the BIC criterion (leaps and bounds or stepwise)

Answer:

```
BIClm = step(lmmodel, direction = c("both"), k = log(n))
```

```
## Start:  AIC=262.13
## siri ~ age + weight + height + adipos + free + neck + chest +
##     abdom + hip + thigh + knee + ankle + biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - hip      1       0.0  487.8 256.61
## - neck     1       0.3  488.1 256.74
## - wrist    1       0.6  488.4 256.88
## - age      1       1.0  488.9 257.08
## - knee     1       3.1  490.9 258.04
## - height   1       3.1  490.9 258.04
## - biceps   1       6.2  494.0 259.45
## - ankle    1       7.9  495.8 260.26
## <none>                  487.8 262.13
## - chest    1      22.0  509.8 266.60
## - forearm  1      24.6  512.4 267.75
## - thigh    1      26.1  514.0 268.45
## - abdom    1      29.7  517.5 270.00
## - adipos   1      41.8  529.6 275.24
## - weight   1     531.0 1018.8 423.77
## - free     1    3347.2 3835.1 724.67
```

```
##
## Step:  AIC=256.61
## siri ~ age + weight + height + adipos + free + neck + chest +
##     abdom + thigh + knee + ankle + biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - neck     1       0.3  488.2 251.23
## - wrist    1       0.6  488.4 251.35
## - age      1       1.0  488.9 251.55
## - knee     1       3.1  491.0 252.52
## - height   1       3.2  491.0 252.54
## - biceps   1       6.2  494.0 253.94
## - ankle    1       7.9  495.8 254.74
## <none>                  487.8 256.61
## - chest    1      23.8  511.7 261.91
## + hip      1       0.0  487.8 262.13
## - forearm  1      24.7  512.6 262.30
## - thigh    1      28.9  516.7 264.13
## - abdom    1      31.1  519.0 265.12
## - adipos   1      42.6  530.5 270.09
## - weight   1     668.3 1156.2 446.95
## - free     1    3389.8 3877.7 721.65
##
## Step:  AIC=251.23
## siri ~ age + weight + height + adipos + free + chest + abdom +
##     thigh + knee + ankle + biceps + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - wrist    1       0.5  488.6 245.91
## - age      1       0.8  489.0 246.09
## - height   1       3.0  491.2 247.10
## - knee     1       3.5  491.7 247.34
## - biceps   1       6.0  494.2 248.48
## - ankle    1       8.2  496.4 249.49
## <none>                  488.2 251.23
## - chest    1      23.9  512.1 256.57
## + neck     1       0.3  487.8 256.61
## + hip      1       0.0  488.1 256.74
## - forearm  1      24.5  512.6 256.81
## - thigh    1      28.8  516.9 258.69
## - abdom    1      30.9  519.1 259.65
## - adipos   1      44.8  532.9 265.62
## - weight   1     670.8 1159.0 441.97
## - free     1    3524.5 4012.6 723.88
##
```

```
## Step:  AIC=245.91
## siri ~ age + weight + height + adipos + free + chest + abdom +
##     thigh + knee + ankle + biceps + forearm
##
##           Df Sum of Sq    RSS    AIC
## - age      1       1.8  490.4 241.22
## - height   1       3.3  492.0 241.92
## - knee     1       3.6  492.2 242.03
## - biceps   1       6.0  494.6 243.16
## - ankle    1       9.7  498.3 244.82
## <none>                  488.6 245.91
## - chest    1      23.6  512.2 251.07
## + wrist    1       0.5  488.2 251.23
## + neck     1       0.2  488.4 251.35
## + hip      1       0.0  488.6 251.42
## - thigh    1      28.5  517.1 253.23
## - forearm  1      30.2  518.9 254.01
## - abdom    1      31.3  519.9 254.45
## - adipos   1      44.3  533.0 260.10
## - weight   1     671.9 1160.6 436.75
## - free     1    3794.7 4283.3 733.17
##
## Step:  AIC=241.22
## siri ~ weight + height + adipos + free + chest + abdom + thigh +
##     knee + ankle + biceps + forearm
##
##           Df Sum of Sq    RSS    AIC
## - height   1       3.1  493.5 237.11
## - knee     1       5.5  495.9 238.20
## - biceps   1       7.0  497.4 238.90
## - ankle    1       9.4  499.8 239.99
## <none>                  490.4 241.22
## + age      1       1.8  488.6 245.91
## + wrist    1       1.4  489.0 246.09
## + hip      1       0.0  490.4 246.74
## + neck     1       0.0  490.4 246.75
## - chest    1      24.9  515.3 246.91
## - thigh    1      27.7  518.1 248.16
## - forearm  1      29.1  519.5 248.75
## - abdom    1      41.3  531.7 254.02
## - adipos   1      44.1  534.5 255.23
## - weight   1     682.4 1172.8 433.61
## - free     1    3794.8 4285.3 727.75
##
## Step:  AIC=237.11
## siri ~ weight + adipos + free + chest + abdom + thigh + knee +
##     ankle + biceps + forearm
##
##           Df Sum of Sq    RSS    AIC
## - knee     1       4.9  498.4 233.81
## - biceps   1       7.3  500.9 234.94
## - ankle    1      10.0  503.5 236.13
## <none>                  493.5 237.11
## + height   1       3.1  490.4 241.22
## + wrist    1       1.8  491.8 241.83
## + age      1       1.6  492.0 241.92
## + hip      1       0.1  493.4 242.60
## + neck     1       0.0  493.5 242.63
## - chest    1      25.0  518.5 242.78
## - thigh    1      25.5  519.0 243.02
## - forearm  1      30.5  524.0 245.18
## - abdom    1      43.1  536.6 250.58
## - adipos   1      77.3  570.8 264.60
## - weight   1     806.1 1299.6 451.38
## - free     1    3810.4 4303.9 723.20
##
## Step:  AIC=233.81
## siri ~ weight + adipos + free + chest + abdom + thigh + ankle +
##     biceps + forearm
##
##           Df Sum of Sq    RSS    AIC
## - biceps   1       6.9  505.3 231.41
## <none>                  498.4 233.81
## - ankle    1      13.9  512.3 234.53
## + knee     1       4.9  493.5 237.11
## + age      1       3.3  495.1 237.85
## + wrist    1       2.5  495.8 238.18
## + height   1       2.5  495.9 238.20
## - chest    1      24.5  522.9 239.18
## + hip      1       0.0  498.4 239.33
## + neck     1       0.0  498.4 239.34
## - forearm  1      32.6  531.0 242.68
## - thigh    1      33.4  531.8 243.02
## - abdom    1      48.7  547.1 249.45
## - adipos   1      91.6  589.9 266.57
## - weight   1     876.9 1375.3 458.70
## - free     1    3808.9 4307.3 717.85
##
## Step:  AIC=231.41
## siri ~ weight + adipos + free + chest + abdom + thigh + ankle +
##     forearm
##
##           Df Sum of Sq    RSS    AIC
## <none>                  505.3 231.41
## - ankle    1      13.0  518.3 231.64
## + biceps   1       6.9  498.4 233.81
## + knee     1       4.4  500.9 234.94
## + age      1       4.3  501.0 234.98
## + wrist    1       3.0  502.3 235.60
## + height   1       2.8  502.5 235.66
## + hip      1       0.2  505.1 236.86
## + neck     1       0.1  505.2 236.91
## - chest    1      26.7  532.0 237.55
## - thigh    1      39.8  545.1 243.08
## - abdom    1      45.2  550.4 245.31
## - forearm  1      49.1  554.4 246.92
## - adipos   1      86.7  592.0 261.82
## - weight   1     909.3 1414.6 459.57
## - free     1    3805.8 4311.1 712.52
```

```
mse(predict(BIClm, newdata = fattest), fattest$siri)
```

```
## [1] 2.349131
```

d) Repeat exercise b) for scaled principal components regression, where you keep enough components to account for 90% of the variation in predictor variables.

Answer:

```
summary(prcomp(fattrain[,-1], scale = TRUE))
```

```
## Importance of components:
##                          PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.091 1.2565 1.02148 0.80898 0.76463 0.57765 0.56200
## Proportion of Variance 0.637 0.1052 0.06956 0.04363 0.03898 0.02225 0.02106
## Cumulative Proportion  0.637 0.7423 0.81184 0.85547 0.89445 0.91670 0.93775
##                            PC8    PC9    PC10   PC11    PC12    PC13   PC14
## Standard deviation     0.51538 0.4330 0.42112 0.3507 0.27785 0.21964 0.1937
## Proportion of Variance 0.01771 0.0125 0.01182 0.0082 0.00515 0.00322 0.0025
## Cumulative Proportion  0.95546 0.9680 0.97978 0.9880 0.99313 0.99635 0.9989
##                           PC15
## Standard deviation     0.13157
## Proportion of Variance 0.00115
## Cumulative Proportion  1.00000
```

As we can see, the cumulative proportion will exceed 90% when we use the first 6 principal components. Therefore, in the following prediction, we will use the first 6 components.

```
library("pls")
```

```
##
## Attaching package: 'pls'

## The following object is masked from 'package:stats':
##
##     loadings
```

```
pc = pcr(siri ~ ., data = fattrain)
pred = predict(pc, fattest, ncomp = 6)
mse(pred, fattest$siri)
```

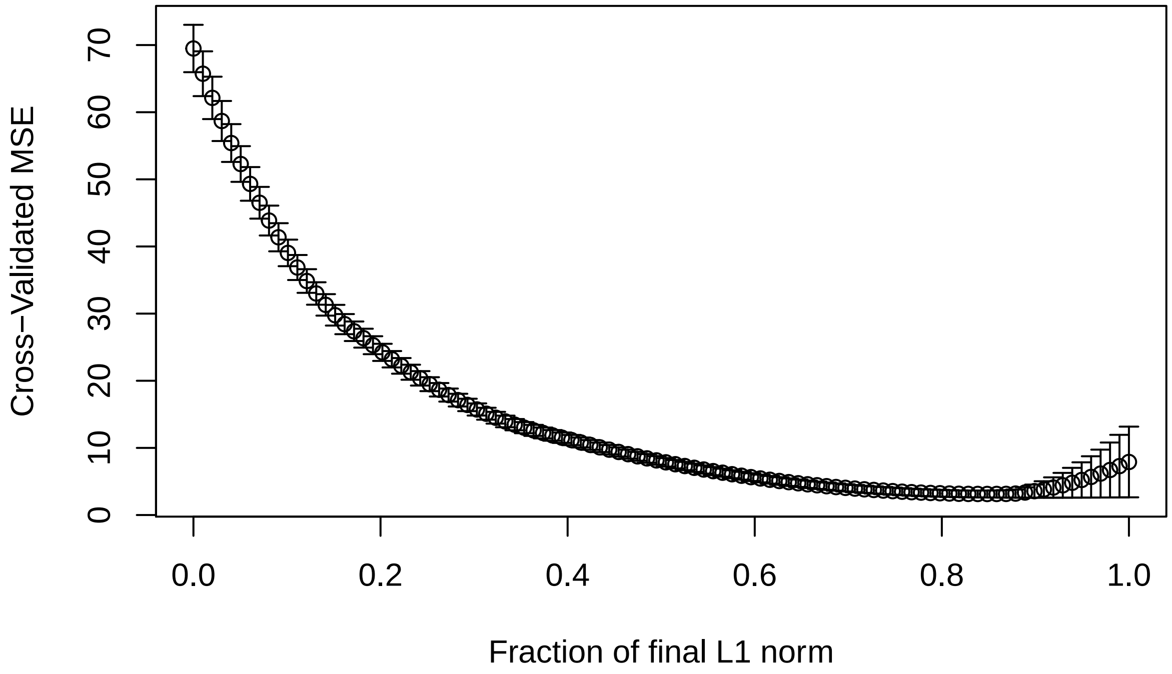
```
## [1] 2.041127
```

e) Repeat exercise b) for Lasso regression, where the amount of shrinkage is selected by 10-fold cross-validation.

```
library("lars")
```

```
## Loaded lars 1.2
```

```
ls = lars(as.matrix(fattrain[,-1]) , fattrain$siri, type = "lasso")
cvml = cv.lars(as.matrix(fattrain[,-1]) , fattrain$siri)
```



```
svm = cvml$index[which.min(cvml$cv)]
predls = predict(ls, fattest[,-1], s = svm, type = "fit", mode = "fraction")$fit
mse(predls, fattest$siri)
```

```
## [1] 2.358614
```

Answer:

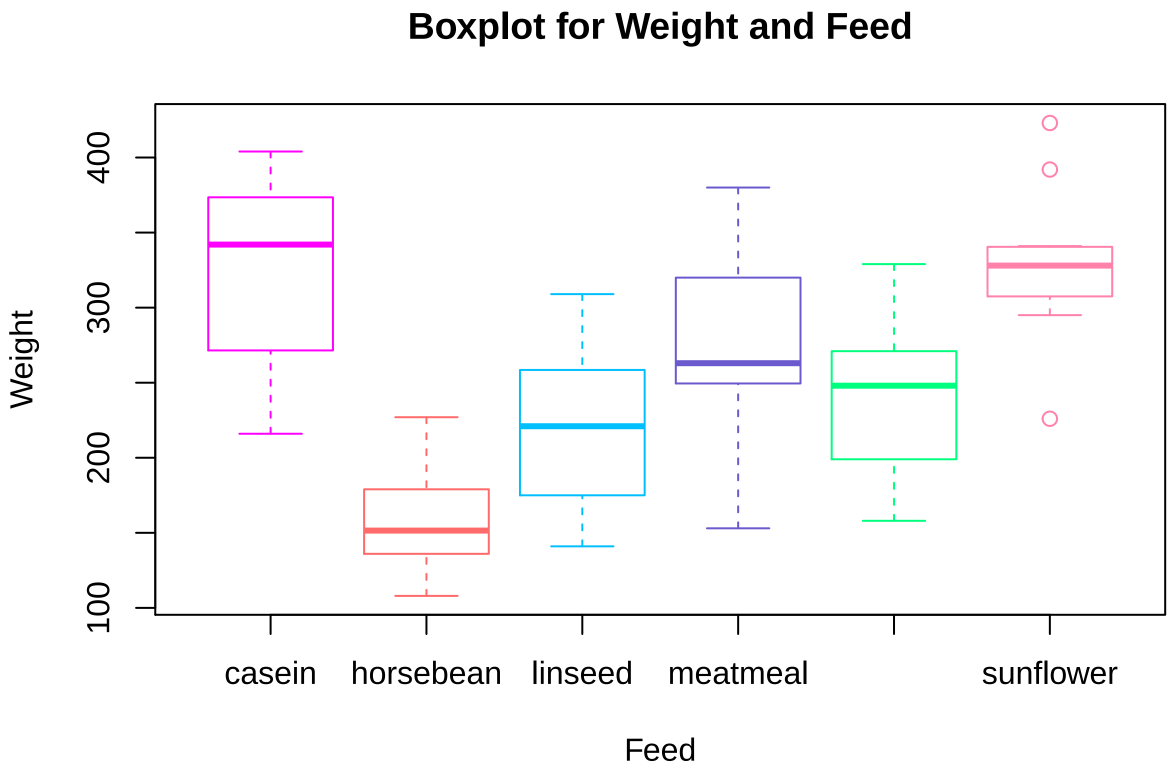
Problem 2

Consider the `chickwts` data in library `datasets`, which compares weights of the chicks randomized into several different groups given different feed supplements.

a) Make boxplots of weight versus feed. Comment on whether the plots show evidence of differences between groups, and whether the data appear consistent with the assumption of normally distributed responses with equal variances.

Answer: We can make a boxplot as below.

```r
cols <- c('magenta', 'indianred1', 'deepskyblue',
          'slateblue', 'springgreen', 'palevioletred1')
boxplot(weight ~ feed, data = chickwts,
        col = 'white',
        border = cols,
        xlab = 'Feed', ylab = 'Weight',
        main = 'Boxplot for Weight and Feed')
```



Note that indeed there are differences between groups since many of the boxes do not overlap and the centers appear to vary a lot. For example, for the feed `casein` and `horsebean`, nearly 90% of their data do not match. The normality with equal variances assumption do not hold since many of the boxes do not have equal inter-quartile ranges; see `sunflower` and `casein` for examples.

b) Perform an F test for equality of treatment means. State the null and alternative hypotheses, and indicate whether there is a significant feed effect at level $\alpha = 0.05$.

Answer: Let the means for equal group be μ_i for $i = 1, \ldots, 6$. The hypotheses are as below

$$H_0\colon\ \mu_1 = \cdots = \mu_6 \quad \text{versus} \quad H_1\colon\ \text{Otherwise}$$

We can perform an F-test as below.

```r
mod <- aov(weight ~ feed, data = chickwts)
summary(mod)
```

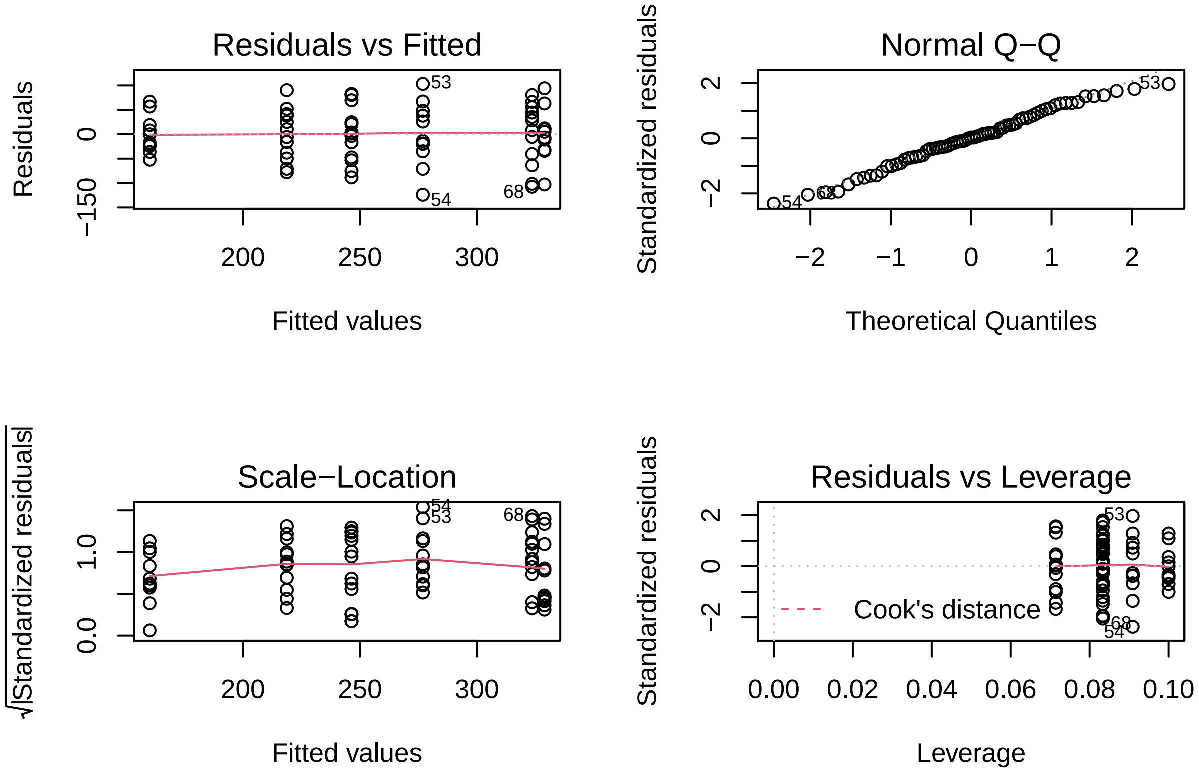
```
##             Df Sum Sq Mean Sq F value   Pr(>F)    
## feed         5 231129   46226   15.37 5.94e-10 ***
## Residuals   65 195556    3009                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Indeed, with a p-value $< .001$, we reject H_0 and conclude that there is a significant feed effects.

c) Check the model assumptions using plots of residuals versus fitted values, QQ plot of standardized residuals versus noraml quantiles, and plot of absolute residuals versus fitted values. Comment on what the plots say about the appropriateness of the assumptions of the F test.

Answer: We can make a diagnostics plot as below.

```r
par(mfrow = c(2, 2))
plot(mod)
```



As we can see from the upperleft plot, the linearity assumption is fine with an almost horizontal line. From the QQ plot, we can see that the normality assumption is fine with points closely scattered around the reference line. For the scale-location plot, we can see some trends with a smoothing line that is horizontal. It implies that the constant variance assumptions do not hold as analyzed in (a).

d) Use the Bonferroni method to test all the pairwise differences between treatment means, controlling the family-wise type I error rate at level $\alpha = 0.05$.

Answer: We can use Bonferroni method to do the test as below.

```
with(chickwts, pairwise.t.test(x = weight, g = feed,
p.adjust.method = 'bonferroni'))
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  weight and feed
##
##           casein  horsebean linseed meatmeal soybean
## horsebean 3.1e-08 -         -       -        -
## linseed   0.00022 0.22833   -       -        -
## meatmeal  0.68350 0.00011   0.20218 -        -
## soybean   0.00998 0.00487   1.00000 1.00000  -
## sunflower 1.00000 1.2e-08   9.3e-05 0.39653  0.00447
##
```

```
## P value adjustment method: bonferroni
```

e) Use the Tukey method to obtain all the pairwise confidence intervals for differences between treatment means, with family-wise confidence level of at least 95%.

Answer: We can use Tukey method to obtain the pairwise confidence intervals as below.

```
TukeyHSD(mod)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = weight ~ feed, data = chickwts)
##
## $feed
##                            diff         lwr       upr     p adj
## horsebean-casein    -163.383333 -232.346876 -94.41979 0.0000000
## linseed-casein      -104.833333 -170.587491 -39.07918 0.0002100
## meatmeal-casein      -46.674242 -113.906207  20.55772 0.3324584
## soybean-casein       -77.154762 -140.517054 -13.79247 0.0083653
## sunflower-casein       5.333333  -60.420825  71.08749 0.9998902
## linseed-horsebean     58.550000  -10.413543 127.51354 0.1413329
## meatmeal-horsebean   116.709091   46.335105 187.08308 0.0001062
## soybean-horsebean     86.228571   19.541684 152.91546 0.0042167
## sunflower-horsebean  168.716667   99.753124 237.68021 0.0000000
## meatmeal-linseed      58.159091   -9.072873 125.39106 0.1276965
## soybean-linseed       27.678571  -35.683721  91.04086 0.7932853
## sunflower-linseed    110.166667   44.412509 175.92082 0.0000884
## soybean-meatmeal     -30.480519  -95.375109  34.41407 0.7391356
## sunflower-meatmeal    52.007576  -15.224388 119.23954 0.2206962
## sunflower-soybean     82.488095   19.125803 145.85039 0.0038845
```

Problem 3:

Consider the `infmort` data in library `faraway`. The data include per capita income, infant mortality per 1000 births, and oil exporter status for 5 regions of the world.

a) Perform a one-way ANOVA with `mortality` as the response and `region` as the predictor. Is the test significant at level 0.05?

Answer:

```
library(faraway)
data(infmort)

fit = lm(mortality ~ region, data = infmort)
anova(fit)
```

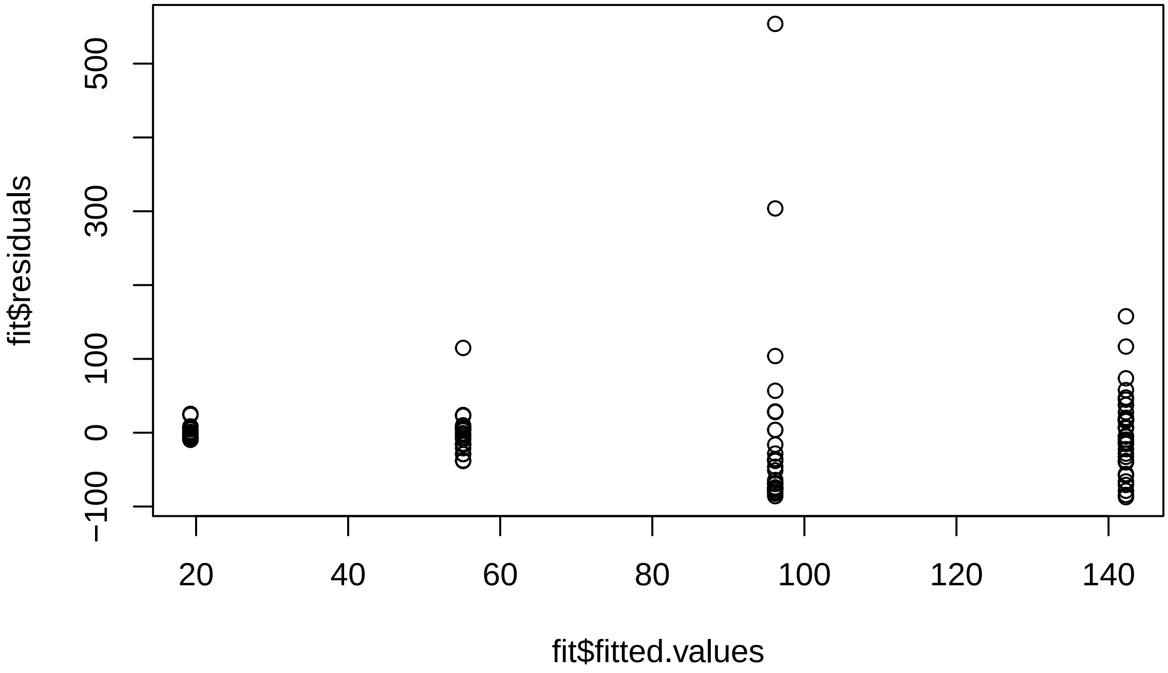
```
## Analysis of Variance Table
##
## Response: mortality
##           Df Sum Sq Mean Sq F value    Pr(>F)
## region     3 210752   70251  11.103 2.494e-06 ***
## Residuals 97 613743    6327
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p-value of 2.494e-06, we have enough evidence to reject the null. We can conclude the mortality rates are not the same across regions.

b) Check the residuals of the model to see if you detect any problems with the model assumptions such as Normal errors with constant variance.

Answer:

```
plot(fit$fitted.values, fit$residuals)
```

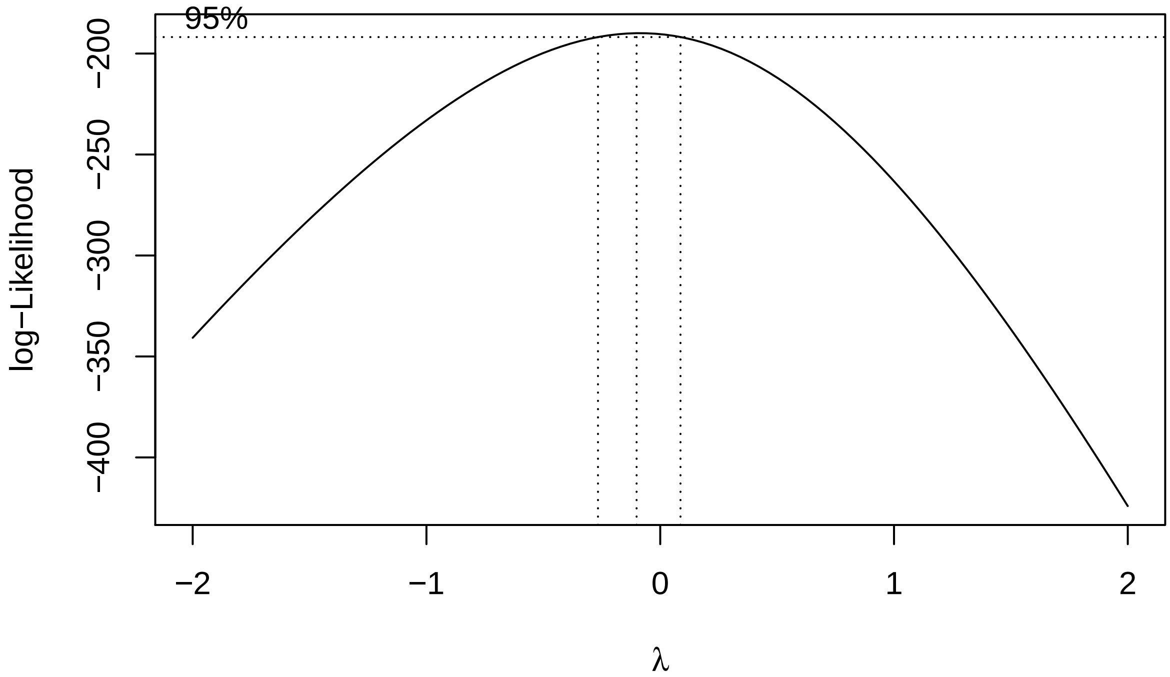


The variance in residuals seems to increase as fitted values increase, so there seems to be evidence of heteroscedasticity. This violates our model assumption.

c) Use the boxcox method to select a transformation of the response. Is the log transformation ($\lambda = 0$) included in the 95% confidence interval for the transformation parameter λ?

Answer:

```
library(MASS)
trans = boxcox(fit)
```



```
lambda = trans$x[trans$y == max(trans$y)]
lambda
```

```
## [1] -0.1010101
```

The log transformation of lambda = 0 is included. The optimal response is lambda = -0.1010101 as shown above.

d) Redo parts a) and b) using log mortality as the response.

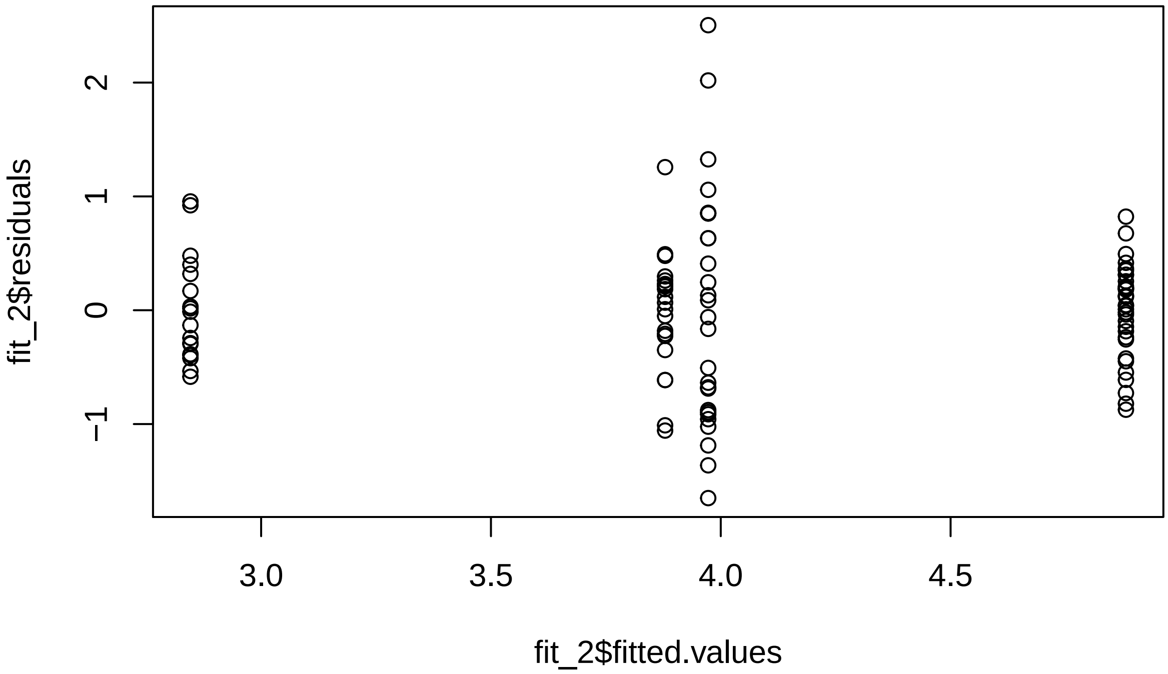
Answer:

```
fit_2 = lm(I(log(mortality)) ~ region, data = infmort)
anova(fit_2)
```

```
## Analysis of Variance Table
##
## Response: I(log(mortality))
##           Df Sum Sq Mean Sq F value    Pr(>F)
## region     3 50.395 16.7985  37.568 3.373e-16 ***
## Residuals 97 43.373  0.4471
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With a p-value of 0, we have enough evidence to reject the null. We can conclude the log mortality rates are not the same across regions.

```
plot(fit_2$fitted.values, fit_2$residuals)
```



Although the variance in residuals is bigger in the middle of the fitted values, the spread as a whole is less cone-shaped and is an improvement from the residual plot of the previous model.

e) With log mortality as the response, which pairs of regions are significantly different, controlling the family-wise type I error rate at 0.05?

Answer:

```
pairwise.t.test(I(log(infmort$mortality)), infmort$region, p.adjust.method = "bonferron
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  I(log(infmort$mortality)) and infmort$region
##
##          Africa  Europe  Asia
## Europe   < 2e-16 -       -
## Asia     4.9e-06 1.6e-06 -
## Americas 2.0e-06 2.7e-05 1
##
## P value adjustment method: bonferroni
```

We can conclude that all pairs of regions are statistically different except Americas vs Asia.

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:

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

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 $\mathbf{X}(\mathbf{X^TX})^{-1}\mathbf{X^T}$.

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

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

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

$$\mathbf{Hy} = \mathbf{X}(\mathbf{X^TX})^{-\mathbf{1}}\mathbf{X^Ty} = \mathbf{X}\hat{\beta}$$

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:

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

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

True:

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

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.

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

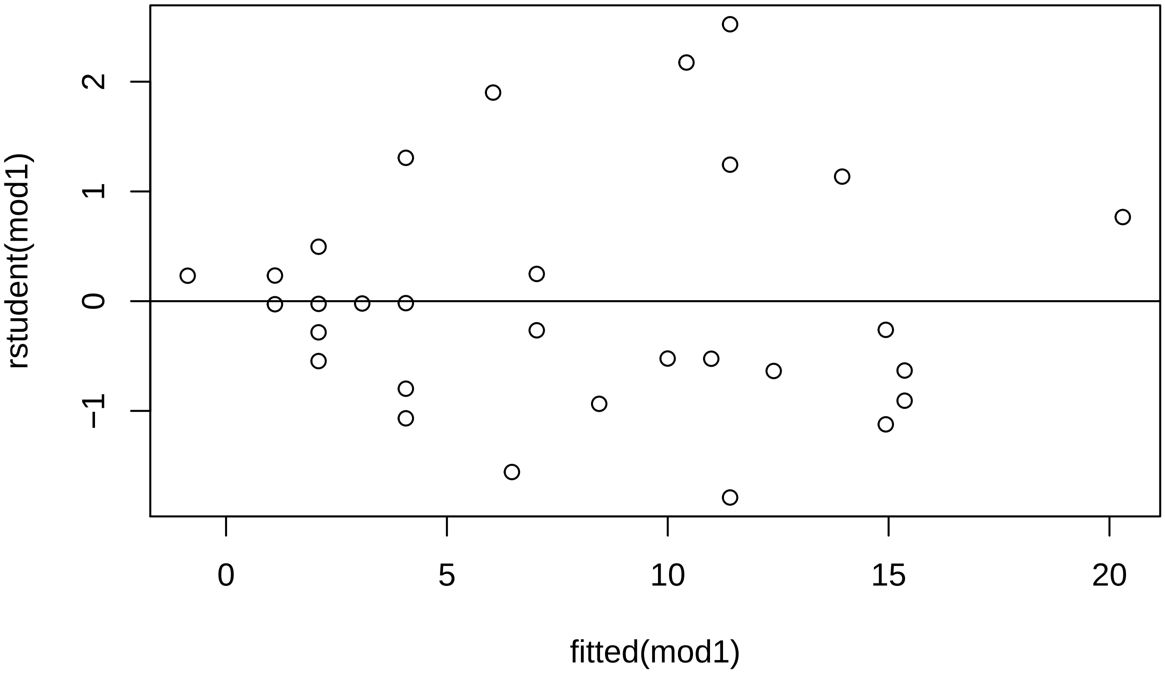
```
## $fit
##        fit      lwr      upr
## 1 11.41096 4.383872 18.43804
##
## $se.fit
## [1] 1.251574
##
## $df
## [1] 27
##
## $residual.scale
## [1] 3.931175
```

$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.

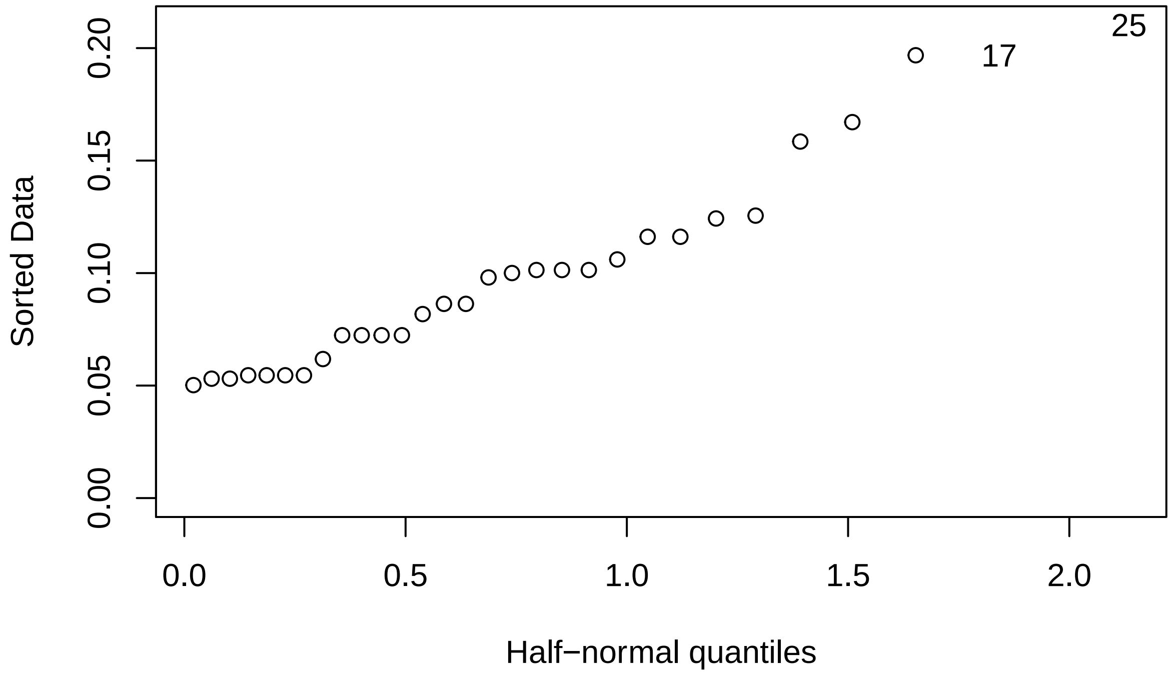


There is a hint of increasing vertical spread in the graph as we move from left to right. If that effect is real, what does it suggest about the model assumptions we have made? Describe briefly.

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?

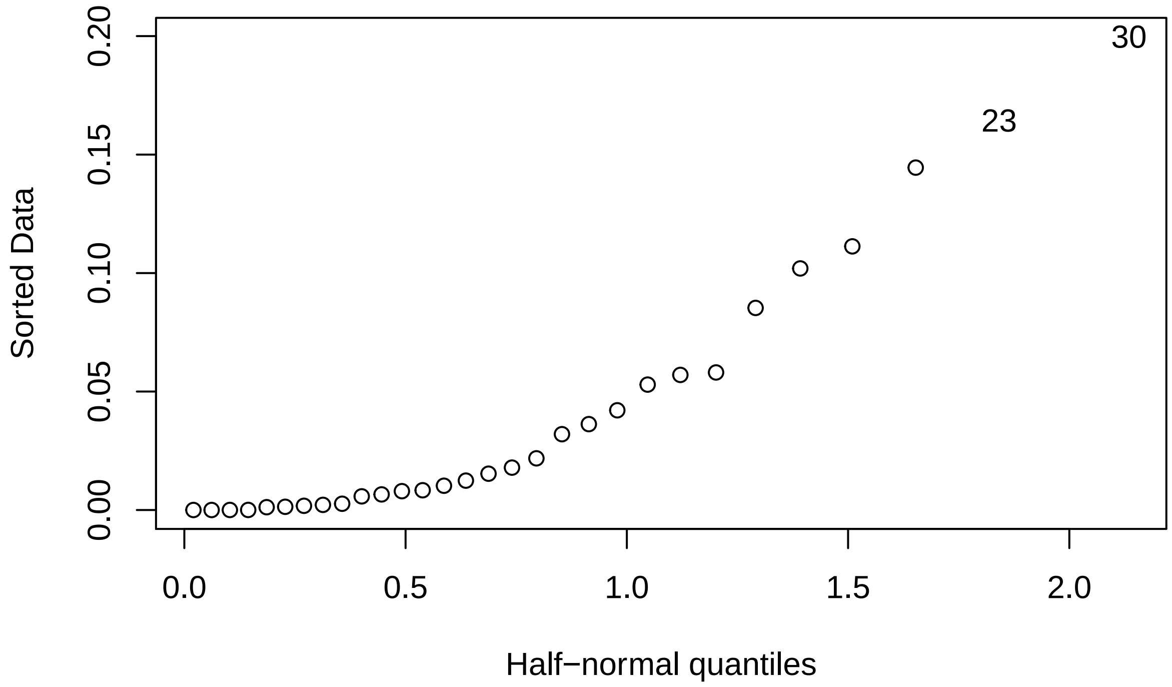
```
library(faraway)
halfnorm(influence(mod1)$hat)
```



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?

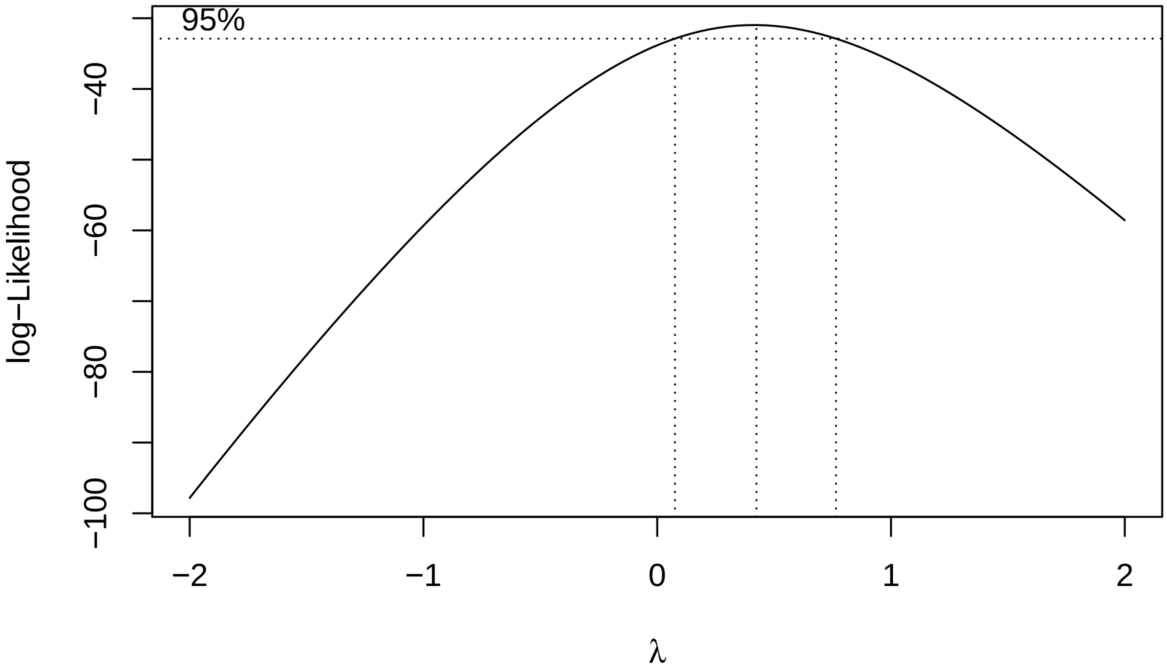
```
halfnorm(cooks.distance(mod1))
```



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).