MAST9014: Introduction to Statistical Learning

Solutions to Week 5 Lab

The first questions use the 'sleep' dataset, which you can download from the course website. This dataset contains (among other things) data on the body weight (kg) and brain weight (g) of 62 mammals. Use the following commands to read the data:

```
mammals <- read.csv("../data/sleep.csv")
mammals$BodyWt <- log(mammals$BodyWt)
mammals$BrainWt <- log(mammals$BrainWt)
```

This creates a data frame, `mammals`, with components (among others) named `BodyWt` and `BrainWt`, then applies a logarithmic transformation to both `BodyWt` and `BrainWt`.

1. Fit a linear model explaining brain weight from body weight, using the `lm` command.

   Display the summary of the fitted model, and then create a scatter plot of the data and superimpose the fitted regression line on it. Does it look like a reasonable fit?

   Use diagnostic plots to assess if the model assumptions are satisfied.

   Solution:

```
model <- lm(BrainWt ~ BodyWt, data = mammals)
summary(model)

##
## Call:
## lm(formula = BrainWt ~ BodyWt, data = mammals)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71550 -0.49228 -0.06162  0.43597  1.94829
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.13479    0.09604   22.23   <2e-16 ***
## BodyWt       0.75169    0.02846   26.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6943 on 60 degrees of freedom
## Multiple R-squared:  0.9208,Adjusted R-squared:  0.9195
## F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16

plot(mammals$BodyWt, mammals$BrainWt)
abline(model, col="red")
```
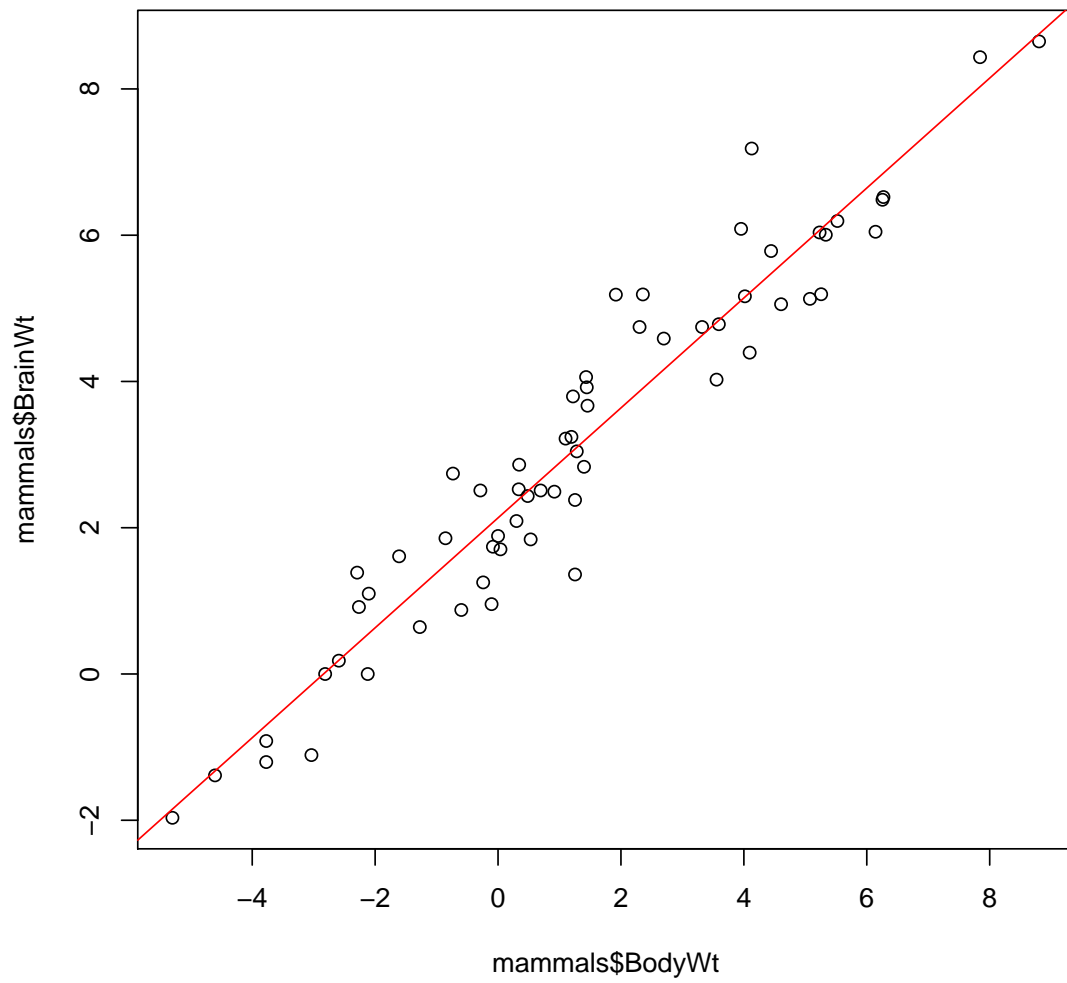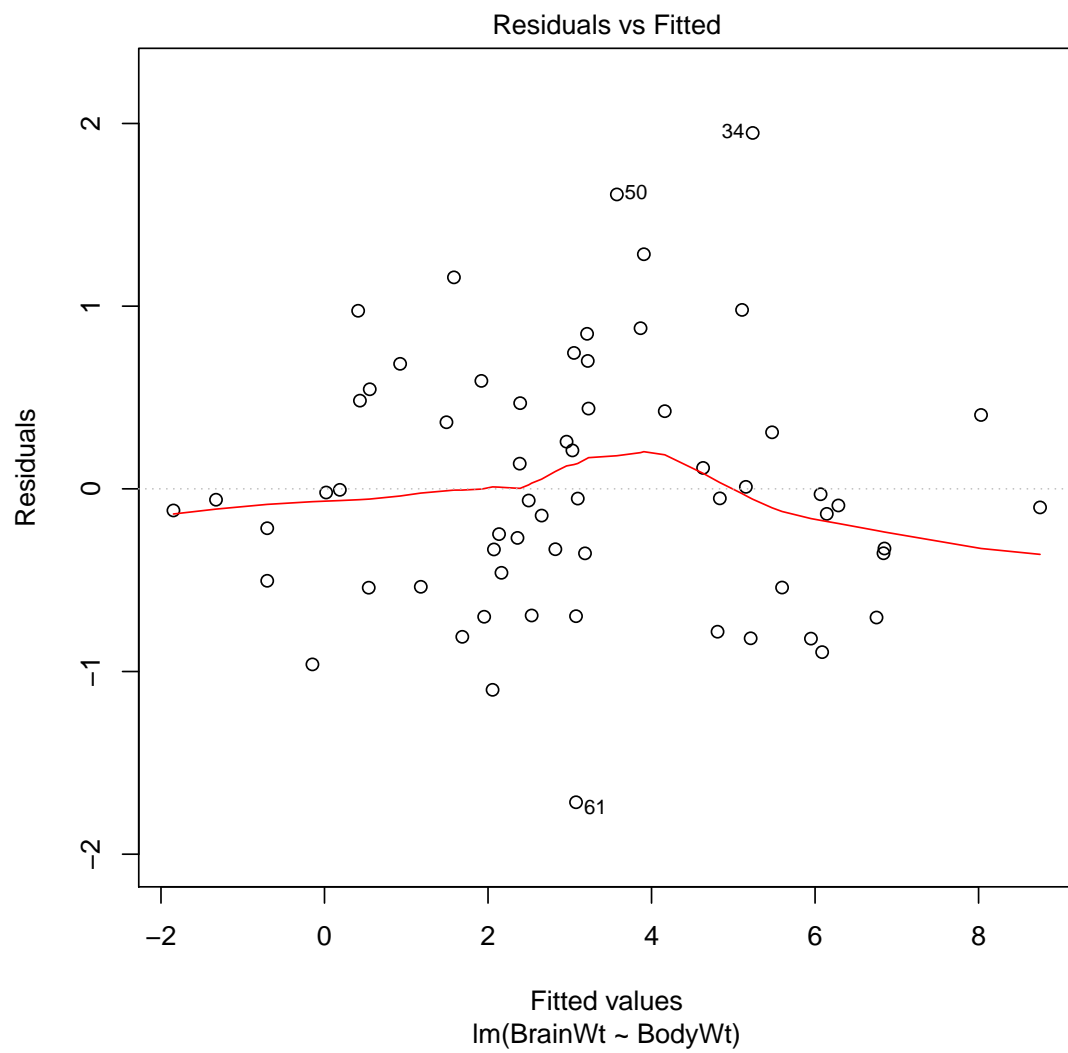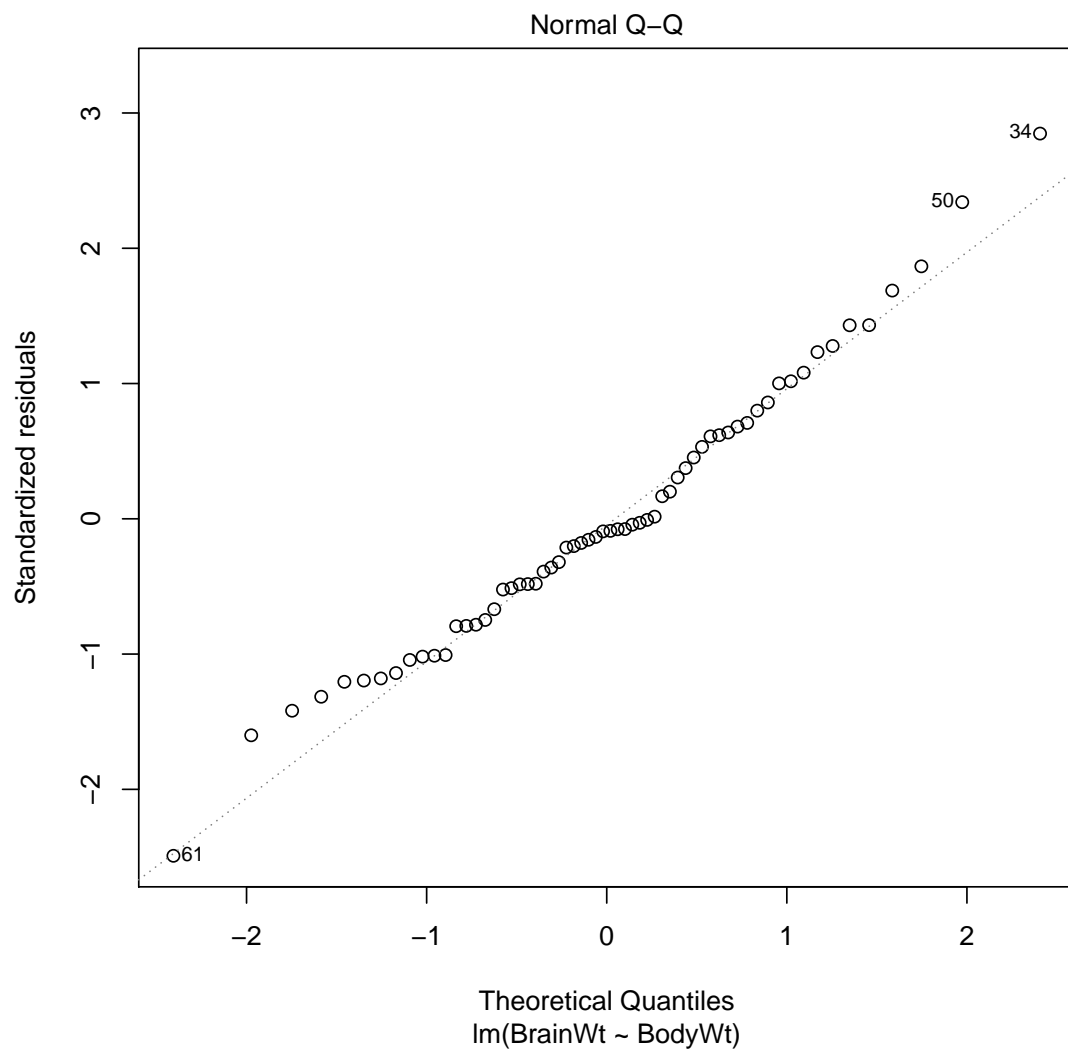
The fit is good.

```
plot(model, which=1)
```
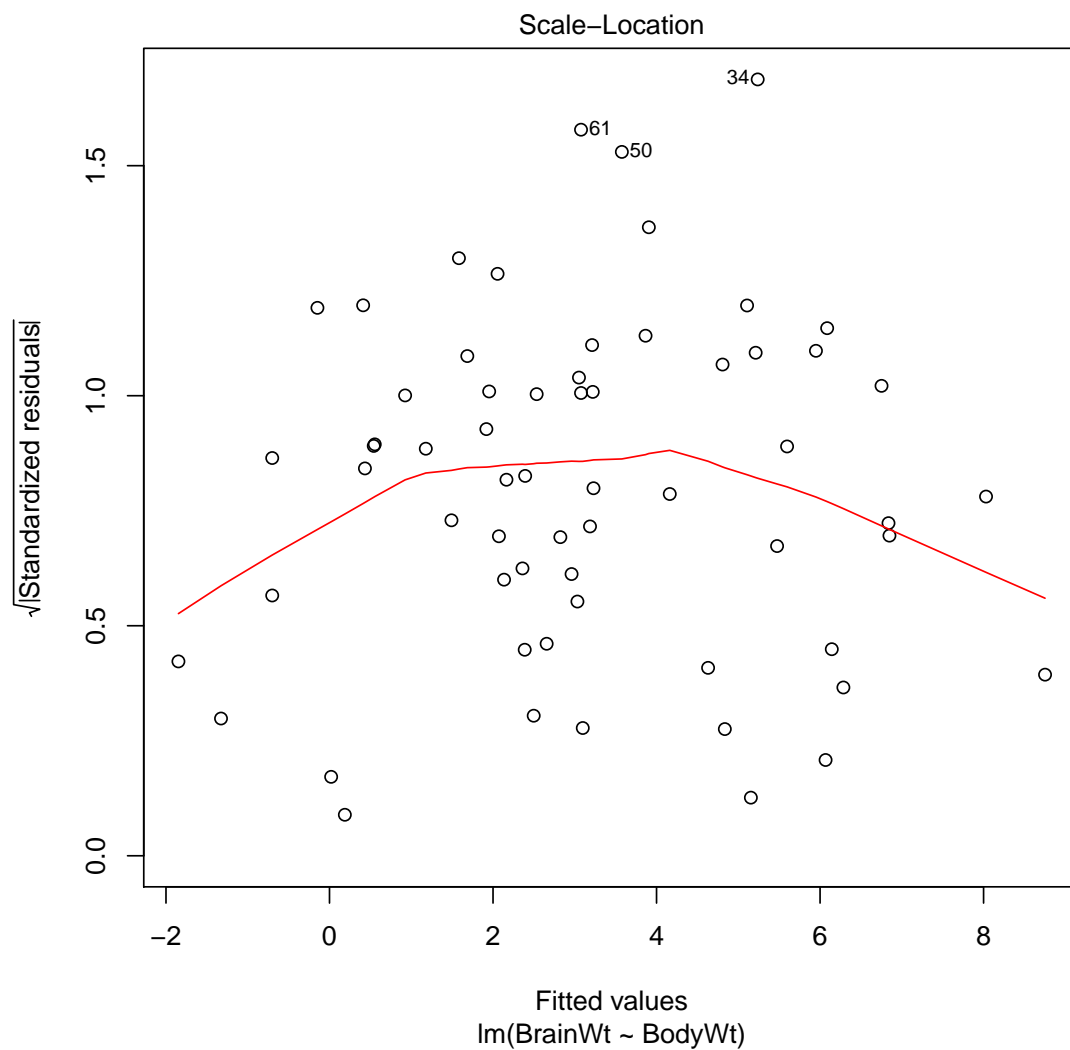
Residuals vs Fitted

lm(BrainWt ~ BodyWt)

The residuals show a slight trend toward negativity as the fitted values increase, but not enough to be a problem.

```
plot(model, which=2)
```

Normal Q–Q

The Q-Q plot looks reasonably linear.

```
plot(model, which=3)
```

Scale–Location

lm(BrainWt ~ BodyWt)

The standardised residuals get smaller on both sides of the plot. This is not ideal, but the lack of a definite trend makes it difficult to correct.

```
plot(model, which=5)
```

Residuals vs Leverage

lm(BrainWt ~ BodyWt)
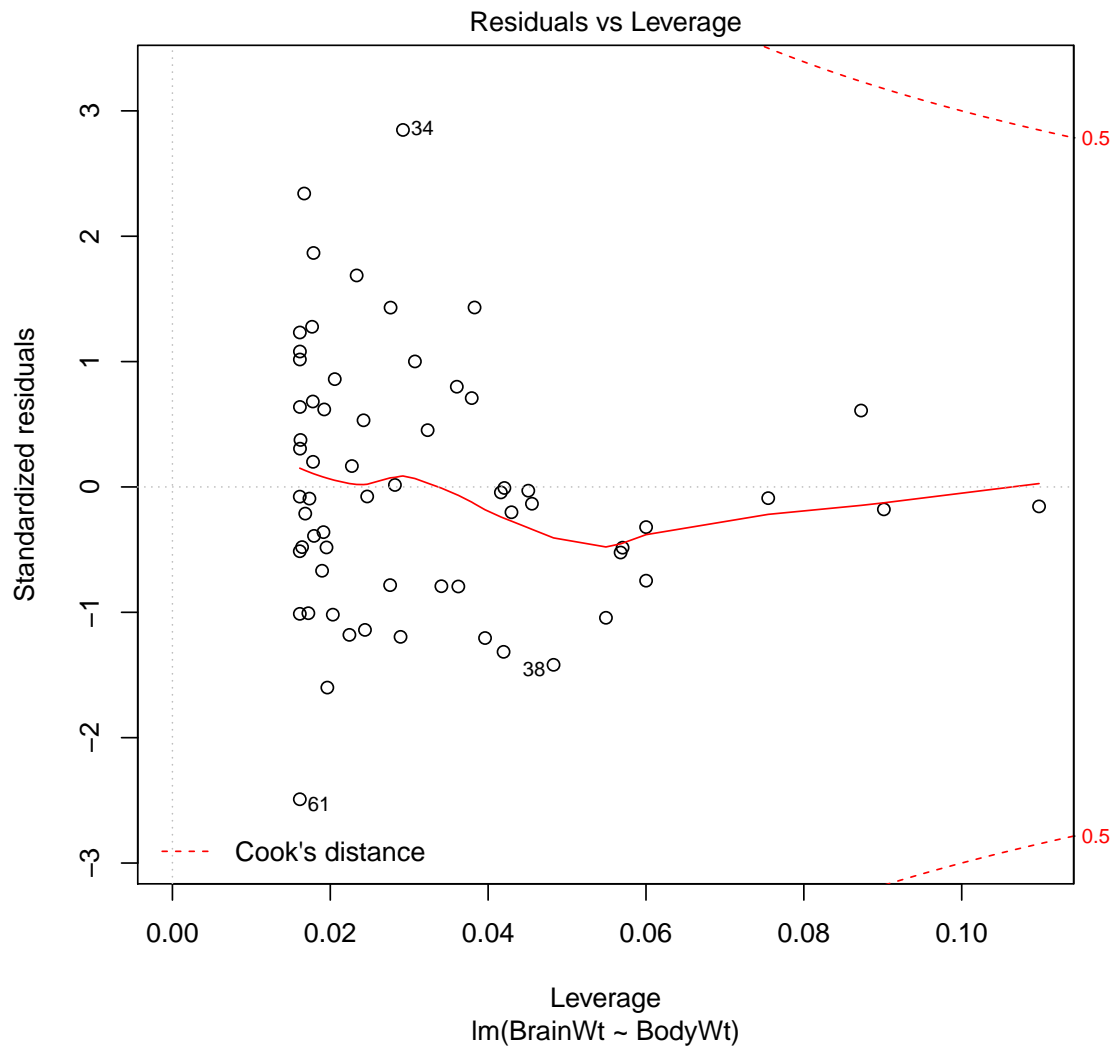
This plot is fine.

2. Using the fitted model or otherwise, obtain:

   (a) The least squares estimator of the parameters, $\mathbf{b}$;

   (b) The vector of residuals, $\mathbf{e}$;

   (c) The residual sum of squares, SS_{Res};

   (d) The regression sum of squares, SS_{Reg};

   (e) The estimator for the variance of the errors, s^2;

   (f) The standardised residuals;

   (g) The leverages of the points; and

   (h) The Cook's distances of the points.

Solution:

```
model$coefficients # parameter estimates

## (Intercept)      BodyWt
##   2.1347887   0.7516859
```

```
str(model$residuals) # first few residuals

##  Named num [1:62] -0.102 -0.248 0.744 -0.332 0.404 ...
##  - attr(*, "names")= chr [1:62] "1" "2" "3" "4" ...

deviance(model) # residual sum of squares

## [1] 28.92271

sum(mammals$BrainWt^2) - deviance(model) # regression ss

## [1] 947.5602

deviance(model)/model$df.residual # sample variance

## [1] 0.4820452

str(rstandard(model)) # standardised residuals

##  Named num [1:62] -0.155 -0.36 1.081 -0.482 0.61 ...
##  - attr(*, "names")= chr [1:62] "1" "2" "3" "4" ...

str(lm.influence(model)$hat) # leverages

##  Named num [1:62] 0.1098 0.0191 0.0162 0.0195 0.0873 ...
##  - attr(*, "names")= chr [1:62] "1" "2" "3" "4" ...

str(cooks.distance(model)) # cook's distances

##  Named num [1:62] 0.00148 0.00127 0.00958 0.00232 0.01777 ...
##  - attr(*, "names")= chr [1:62] "1" "2" "3" "4" ...
```

3. Find a 95% confidence interval for a mammal weighing 50 kg.
   Solution:

```
predict(model, data.frame(BodyWt = log(50)), interval = "confidence", level = 0.95)

##       fit      lwr      upr
## 1 5.075401 4.846066 5.304736
```

4. Find a 95% prediction interval for a mammal weighing 50 kg.
   Solution:

```
predict(model, data.frame(BodyWt = log(50)), interval = "prediction", level = 0.95)

##       fit      lwr      upr
## 1 5.075401 3.667797 6.483006
```

5. Test the following hypotheses, using the **anova** function.

   (a) $H_0 : \boldsymbol{\beta} = 0$
   (b) $H_0 : \beta_1 = 0$

7

(c) $H_0 : \beta_0 = 0$

(d) $H_0 : \boldsymbol{\beta} = (2, 1)$

Solution:

```
null <- lm(BrainWt ~ 0, data = mammals)
anova(null, model)

## Analysis of Variance Table
##
## Model 1: BrainWt ~ 0
## Model 2: BrainWt ~ BodyWt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     62 976.48
## 2     60  28.92  2    947.56 982.85 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

null <- lm(BrainWt ~ 1, data = mammals)
anova(null, model)

## Analysis of Variance Table
##
## Model 1: BrainWt ~ 1
## Model 2: BrainWt ~ BodyWt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     61 365.11
## 2     60  28.92  1    336.19 697.42 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

null <- lm(BrainWt ~ 0 + BodyWt, data = mammals)
anova(null, model)

## Analysis of Variance Table
##
## Model 1: BrainWt ~ 0 + BodyWt
## Model 2: BrainWt ~ BodyWt
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     61 267.079
## 2     60  28.923  1    238.16 494.05 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

library(car)
linearHypothesis(model,diag(2),c(2,1))

## Linear hypothesis test
##
## Hypothesis:
## (Intercept) = 2
## BodyWt = 1
##
## Model 1: restricted model
## Model 2: BrainWt ~ BodyWt
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
```

```
## 1       62 68.024
## 2       60 28.923  2     39.101 40.558 7.199e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
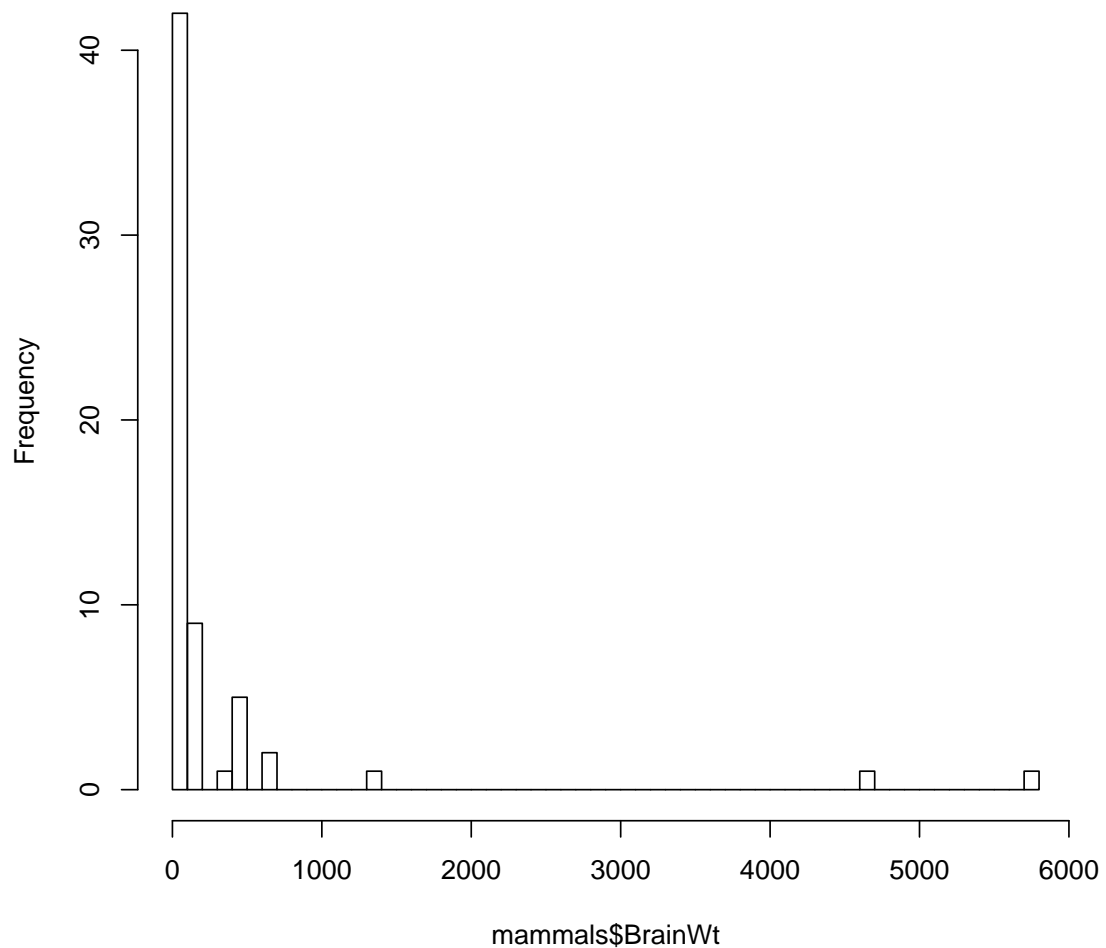
We reject all null hypotheses.

6. By visualising the raw data, justify the use of a double logarithmic transformation. Write down the final model for the (untransformed) brain weight vs. body weight.
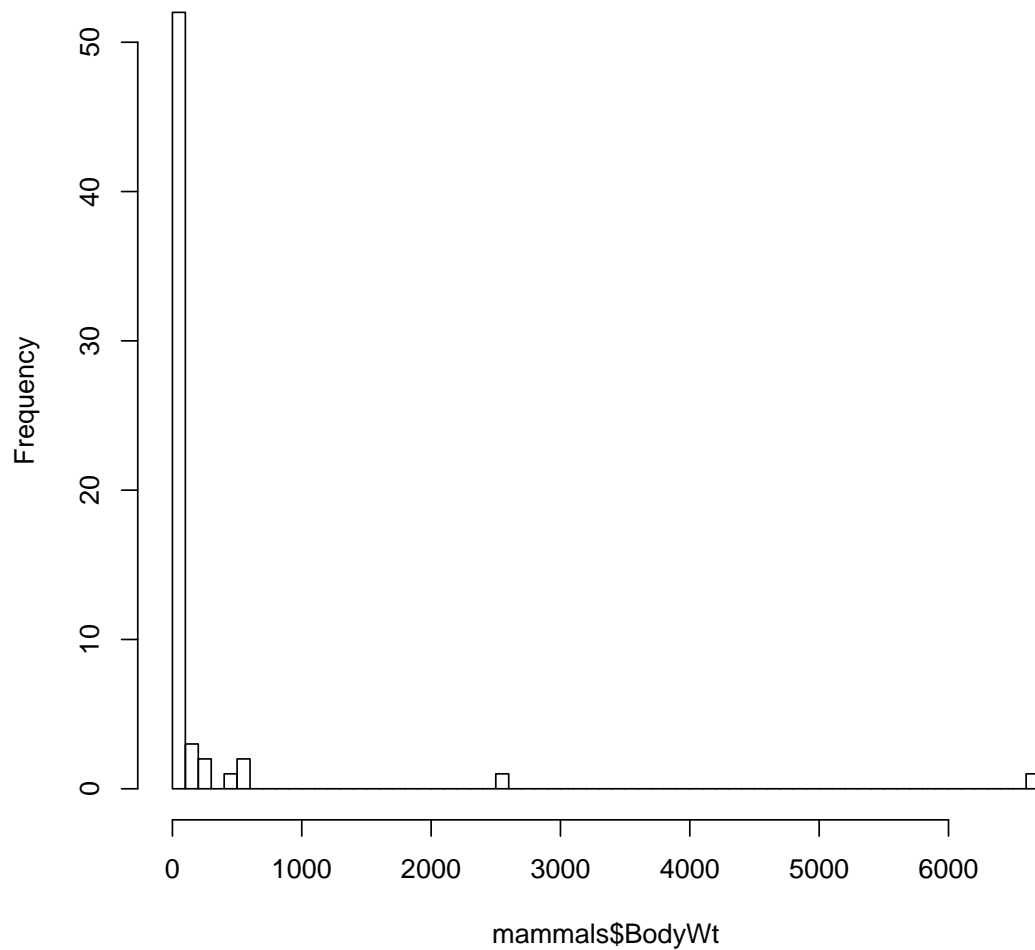
   Solution:

```
mammals <- read.csv('../data/sleep.csv')
hist(mammals$BrainWt,breaks=50)
```

Histogram of mammals$BrainWt



```
hist(mammals$BodyWt,breaks=50)
```

Histogram of mammals$BodyWt

We see that both brain weight and body weight have extremely right-skewed distributions. This is one of the hallmarks of data which requires a logarithmic transformation, particularly in the response (to achieve normal errors). In addition, both variables are constrained to be positive, another indication that a logarithmic transformation may be required.

Merely being right-skewed would not be a strong enough case to transform the predictor, although the extreme nature of the skew results in some points with extremely high leverage/Cook's distance. However, transforming the brain weight alone does not result in a linear relationship, while transforming both brain and body weight results in an obviously linear relationship.

```
plot(log(BrainWt)~BodyWt,data=mammals,xlim=c(0,500))
```
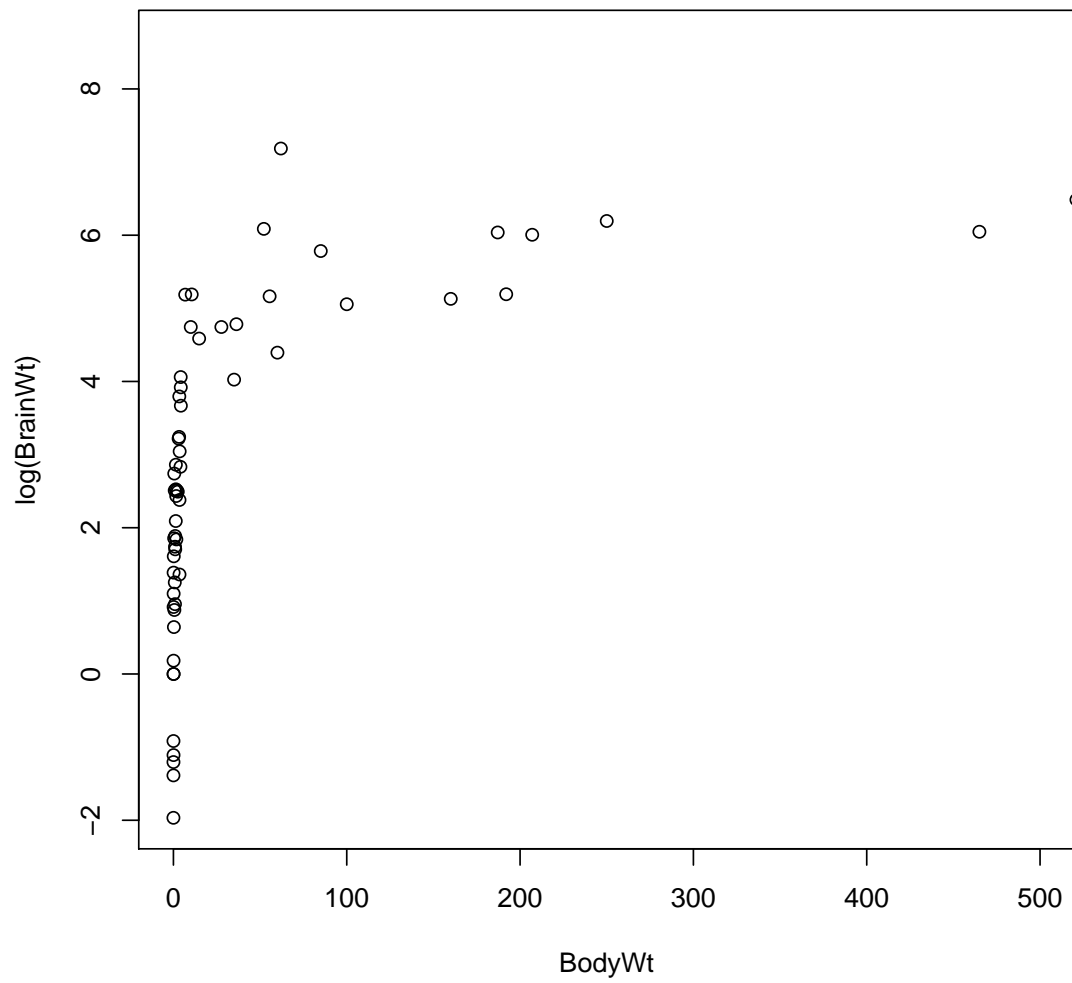
```
plot(log(BrainWt)~log(BodyWt),data=mammals)
```
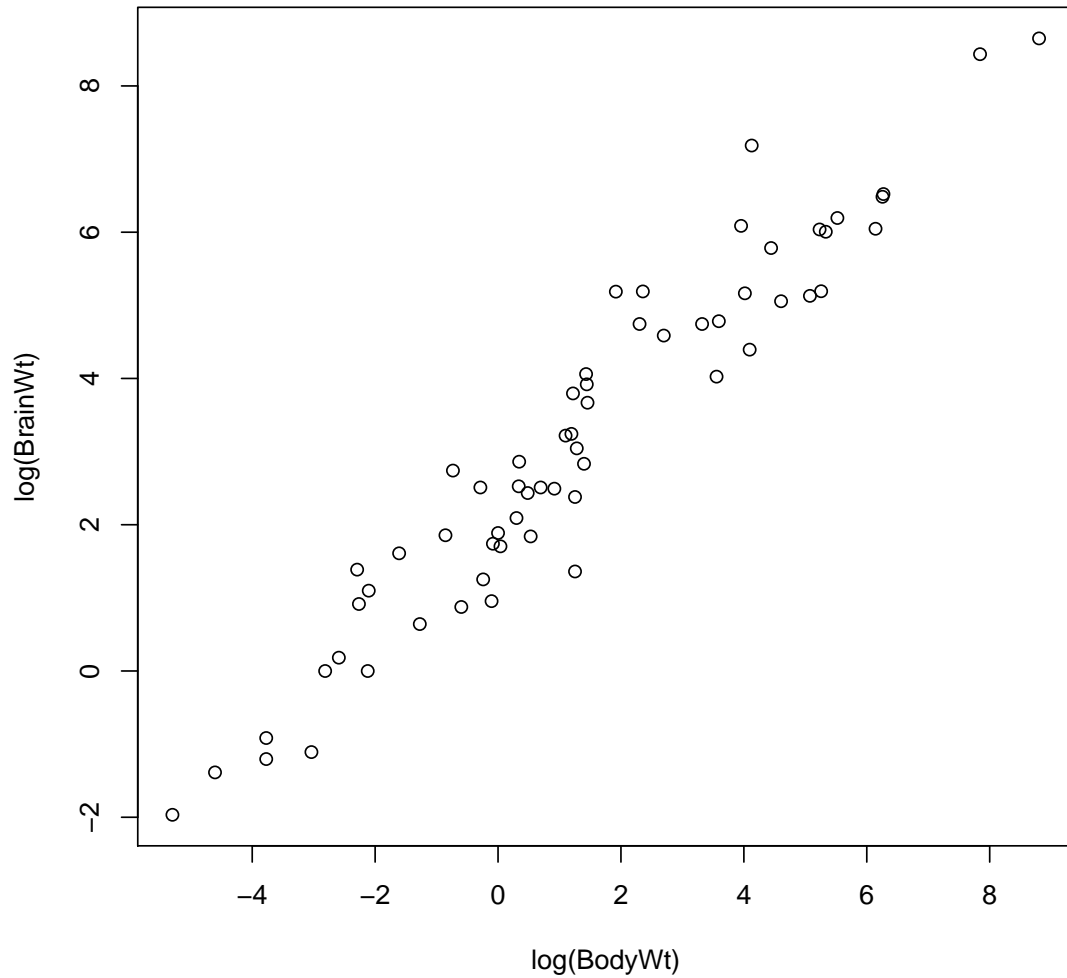
The final model is

$$\text{brain weight} = 8.46 \cdot (\text{body weight})^{0.75} \cdot \varepsilon.$$

7. Suppose X is $n \times p, p \leq n$ of full rank and C is $r \times p, r \leq p$ also of full rank.

   (a) Show that $X^T X$ is positive definite (hint: use the definition).

   Solution: Suppose β is a $p \times 1$ vector. Then $\beta^T(X^T X)\beta = (X\beta)^T(X\beta)$ wich is a sum of squares of the $n \times 1$ vector $X\beta$. Hence, it is non-negative showing that $X^T X$ is positive semidefinite. To show it is positive definite, note that $\beta^T(X^T X)\beta = 0$, implies $X\beta = 0$, wich in turn implies that $X^T X\beta = 0$. But $X^T X$ is rank p so its columns are linearly independent, and hence $\beta = 0$.

   (b) Show that $C(X^T X)^{-1}C^T$ is positive definite (hint: why does $(X^T X)^{-1}$ have a matrix square root?).

   Solution: Since $(X^T X)$ is positive definite and symmetric, Section 1.7 of Module 3 shows that it has a positive definite square root. The square root is of the form $P\Lambda^{1/2}P^T$ where P is orthonormal and the diagonal matrix Λ has the strictly positive eigenvalues(Theorem 2.7) of $X^T X$ on the diagonal. Therefore $(X^T X)^{-1}$ has the positive definite and invertible square root $Q = P\Lambda^{-1/2}P^T$. Hence, since Q is symmetric, $C(X^T X)^{-1}C^T = CQQ^TC^T = Z^T Z$ where $Z = Q^T C^T$. Since the rank of a matrix is unaltered by premultiplication by a non-singular matrix, $r(Z) = r(Q^T C^T) = r(C^T) = r(C) = r$, so Z is full rank. By the previous part of this question applied to Z instead of X, $C(X^T X)^{-1}C^T$ is positive definite.

(c) Show that $C(X^TX)^{-1}C^T$ is invertible.

Solution: From facts about rank, $r(C(X^TX)^{-1}C^T) = r(Z^TZ) = r(Z) = r$ from the previous part of this question.

(d) Show that $[C(X^TX)^{-1}C^T]^{-1}$ is positive definite.

Solution: Since $(X^TX)^{-1}$ is symmetricc, $[C(X^TX)^{-1}C^T]^T = C(X^TX)^{-1}C^T$ is symmetric. The matrix $C(X^TX)^{-1}C^T$ is also positive definite from this question, so it has strictly positive eigenvalues. Hence, the matrix can be expressed as $R\Theta R^T$ where R is an orthogonal matrix of eigenvectors of $C(X^TX)^{-1}C^T$ and, by Theorem 2.7, Θ is a diagonal matrix with strictly positive eigenvalues on the diagonal. Then $R\Theta^{-1}R^T$ is the inverse of $C(X^TX)^{-1}C^T$ and, for any β, $\beta^T R\Theta^{-1}R^T\beta = (R^T\beta)^T\Theta R^T\beta$ is a sum of squares, some of whose entries must be positive because they are elements of the diagonal of Θ multiplied by elements of $(R^T\beta)^T R^T\beta = \beta^T\beta > 0$.

8. In this question we consider the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^*$. The test statistic for this hypothesis is

$$\frac{(\mathbf{b} - \boldsymbol{\beta}^*)^T X^T X(\mathbf{b} - \boldsymbol{\beta}^*)/p}{SS_{Res}/(n-p)}.$$

(a) Show that

$$(\mathbf{b} - \boldsymbol{\beta}^*)^T X^T X(\mathbf{b} - \boldsymbol{\beta}^*) = (\mathbf{y} - X\boldsymbol{\beta}^*)^T(\mathbf{y} - X\boldsymbol{\beta}^*) - (\mathbf{y} - X\mathbf{b})^T(\mathbf{y} - X\mathbf{b}).$$

That is, it is the SS_{Res} for the null model minus the SS_{Res} for the full model.

Also show that, in general,

$$(\mathbf{b} - \boldsymbol{\beta}^*)^T X^T X(\mathbf{b} - \boldsymbol{\beta}^*) \neq \mathbf{y}^T X(X^TX)^{-1}X^T\mathbf{y} - \boldsymbol{\beta}^{*T}X^TX\boldsymbol{\beta}^*.$$

That is, in this case we can not write it as the SS_{Reg} for the full model minus the SS_{Reg} for the model under H_0.

Solution: Letting $H = X(X^T)^{-1}X^T$, equation (1) on p.11 of Module 4 gives

$$\begin{aligned}
(\mathbf{y} - X\boldsymbol{\beta}^*)^T(\mathbf{y} - X\boldsymbol{\beta}^*) & \\
= & (\mathbf{y} - X\mathbf{b})^T(\mathbf{y} - X\mathbf{b}) + (X\mathbf{b} - X\boldsymbol{\beta}^*)^T(X\mathbf{b} - X\boldsymbol{\beta}^*) \\
= & (\mathbf{y} - X\mathbf{b})^T(\mathbf{y} - X\mathbf{b}) + (\mathbf{b} - \boldsymbol{\beta}^*)^T X^T X(\mathbf{b} - \boldsymbol{\beta}^*).
\end{aligned}$$

and rearraanging gives the required equality. Note also that $X\mathbf{b} = H\mathbf{y}$ and H is idempotent so

$$\begin{aligned}
(\mathbf{b} - \boldsymbol{\beta}^*)^T X^T X(\mathbf{b} - \boldsymbol{\beta}^*) & = (X\mathbf{b})^T(X\mathbf{b}) + (X\boldsymbol{\beta}^*)^T(X\boldsymbol{\beta}^*) - 2(X\mathbf{b})^T X\boldsymbol{\beta}^* \\
& = (H\mathbf{y})^T H\mathbf{y} + \boldsymbol{\beta}^{*T}X^TX\boldsymbol{\beta}^* - 2(X\mathbf{b})^T X\boldsymbol{\beta}^* \\
& = \mathbf{y}^T H\mathbf{y} + \boldsymbol{\beta}^{*T}X^TX\boldsymbol{\beta}^* - 2(X\mathbf{b})^T X\boldsymbol{\beta}^* \\
& = y^T H\mathbf{y} - \boldsymbol{\beta}^{*T}X^TX\boldsymbol{\beta}^*
\end{aligned}$$

only if $-2(X\mathbf{b})^T X\boldsymbol{\beta}^* = -2\boldsymbol{\beta}^{*T}X^TX\boldsymbol{\beta}^*$ which, in general, only occurs if $\boldsymbol{\beta}^* = 0$.

(b) Show directly that $(\mathbf{b} - \boldsymbol{\beta}^*)^T X^T X(\mathbf{b} - \boldsymbol{\beta}^*)$ and SS_{Res} are independent, that is without using our existing results that $\mathbf{b}$ and SS_{Res} are independent.

Hint: set $\mathbf{q} = \mathbf{y} - X\boldsymbol{\beta}^*$ then

i. Show that $(\mathbf{b} - \boldsymbol{\beta}^*)^T X^T X(\mathbf{b} - \boldsymbol{\beta}^*) = \mathbf{q}^T X(X^TX)^{-1}X^T\mathbf{q}$.

ii. Show that $SS_{Res} = \mathbf{q}^T[I - X(X^TX)^{-1}X^T]\mathbf{q}$ and hence that these two quadratic forms are independent.

Solution: We express both quantities as quadratic forms in $\mathbf{q}$. Arguing as in the last part,

$$\begin{aligned}
\mathbf{q}^T X(X^TX)^{-1}X^T\mathbf{q} & = \mathbf{y}^T HT\mathbf{y} - 2(X\boldsymbol{\beta}^*)^T H + (X\boldsymbol{\beta}^*)^T HX\boldsymbol{\beta}^* \\
& = (X\mathbf{b})^T X\mathbf{b} - 2(X\boldsymbol{\beta}^*)^T H + (\boldsymbol{\beta}^*)^T X^T X\boldsymbol{\beta}^* \\
& = (\mathbf{b} - \boldsymbol{\beta}^*)^T X^T X(\mathbf{b} - \boldsymbol{\beta}^*).
\end{aligned}$$

For the SS_{Res} note first that

$$\begin{aligned}
\boldsymbol{\beta}^{*T}X^T[I-H]X\boldsymbol{\beta}^* &= \boldsymbol{\beta}^{*T}X^TX\boldsymbol{\beta}^* - \boldsymbol{\beta}^{*T}X^TX(X^TX)^{-1}X^TX\boldsymbol{\beta}^* \\
&= \boldsymbol{\beta}^{*T}X^TX\boldsymbol{\beta}^* - \boldsymbol{\beta}^{*T}X^TX\boldsymbol{\beta}^* \\
&= \mathbf{0}.
\end{aligned}$$

Similarly $\mathbf{y}^T[I-HT]X\boldsymbol{\beta}^* = \mathbf{0}$ and $\boldsymbol{\beta}^{*T}X^T[I-H]\mathbf{y} = \mathbf{0}$, so

$$\begin{aligned}
\mathbf{q}^T[I-H]\mathbf{q} \\
&= \mathbf{y}^T[I-H]\mathbf{y} - \boldsymbol{\beta}^{*T}X^T[I-H]\mathbf{y} \\
&\quad -\mathbf{y}^T[I-H]X\boldsymbol{\beta}^* + \boldsymbol{\beta}^{*T}X^T[I-H]X\boldsymbol{\beta}^* \\
&= \mathbf{y}^T[I-H]\mathbf{y} \\
&= SS_{Res}.
\end{aligned}$$

Finally, we know that var $\mathbf{q} = \sigma^2 I$, so, using our theorem for the independence of quandratic forms

$$\begin{aligned}
AVB &= X(X^TX)^{-1}X^T\sigma^2 I[I-H] \\
&= \sigma^2\left(X(X^TX)^{-1}X^T - X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T\right) \\
&= \sigma^2\left(X(X^TX)^{-1}X^T - X(X^TX)^{-1}X^T\right) \\
&= \mathbf{0}
\end{aligned}$$

as required.

9. Recall the joint confidence region for the parameters of a full rank linear model:

$$(\mathbf{b}-\boldsymbol{\beta})^TX^TX(\mathbf{b}-\boldsymbol{\beta}) \le ps^2 f_\alpha.$$

Use this to derive a test for the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^*$. Show that this test is equivalent to the test for $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^*$ obtained using the general linear hypothesis.

Solution: We do not reject H_0 if and only if $\boldsymbol{\beta}^*$ lies in the joint confidence region, i.e., if and only if

$$(\mathbf{b}-\boldsymbol{\beta}^*)^TX^TX(\mathbf{b}-\boldsymbol{\beta}^*)^T \le ps^2 f_\alpha.$$

But the general linear hypothesis test does not reject H_0 if and only if

$$\begin{aligned}
\frac{(\mathbf{b}-\boldsymbol{\beta}^*)^TX^TX(\mathbf{b}-\boldsymbol{\beta}^*)/p}{SS_{Res}/(n-p)} &\le f_\alpha \\
(\mathbf{b}-\boldsymbol{\beta}^*)^TX^TX(\mathbf{b}-\boldsymbol{\beta}^*) &\le p\frac{SS_{Res}}{n-p}f_\alpha \\
(\mathbf{b}-\boldsymbol{\beta}^*)^TX^TX(\mathbf{b}-\boldsymbol{\beta}^*) &\le ps^2 f_\alpha.
\end{aligned}$$

Therefore the two tests are equivalent.