Section1

Joshua Bean, Miriam Slattery

24/05/2018

The First step to be taken was to enter the data into R to allow us to perform analysis on it.

```
Child<- c(1:12)
Height<-c(108.7,161.29,95.25,100.33,115.57,97.79,109.22,57.15,93.98,59.69,83.82,147.32)
Weight<-c(18.14,42.41,16.10,13.61,23.59,7.71,17.46,3.86,14.97,4.31,9.53,35.83)
Length<-c(37,49.5,34.5,36,43,28,37,20,33.5,30.5,38.5,47.0)
catheter <-data.frame(Child, Height, Weight, Length)
```

Now we have entered the data, we still have no idea of the relationships between the variables. It may be interesting to know the relationships between the predictor variables and the response variable. To do so, we will create a pairwise scatter plot matrix as follows,

```
pairs(subset(catheter, select=c(2:4)))
```

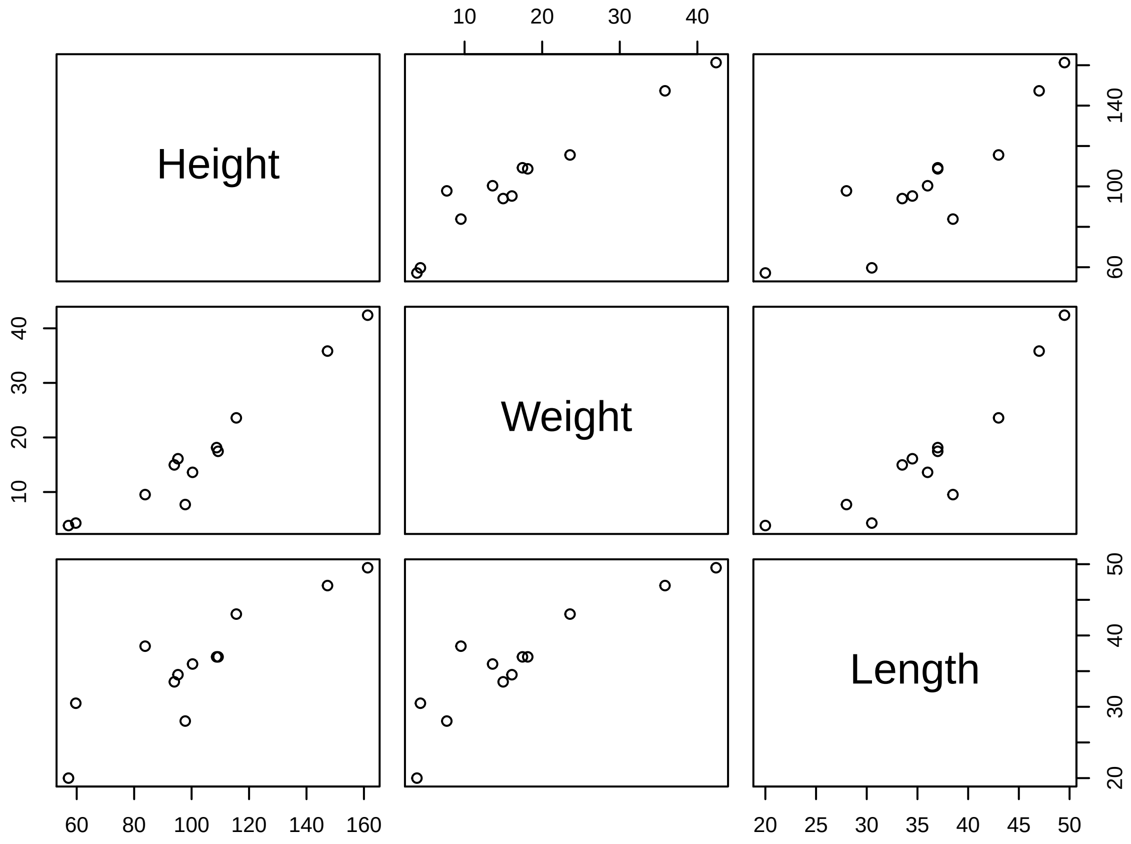

Figure 1: A pariwise scatter plot of all variables.

Before we discuss relationships, it is important to note that there are only 12 data points so there will be a lot of unexplained scatter. To investigate the relationship between Length, the response variable and the two predictor variables, Heigth and Weight, we need to consider the two plots in the top right of Figure 1.

First the scatter plot between Length and Weight (top center plot) shows the relationship is strong, positive and linear. Similarly, the plot between Length and Height (top right) shows the relationship is moderate, positive and mostly linear with slight positive curvature. The trend is only moderate to moderately strong because for the lower values, there is some deviation from the strong trendline. So we can conclude that the relationship between the resonse and the predictor variables is positive linear.

We may also wish to investigate the relationship between the predictor variables, as this will be important later on. To do so, consider the plot of Height against Weight (center rigth plot), which shows the relationship is moderately strong, positive and mostly linear with slight positive curvature, with few outlying points below the curve. Althought give the small number of data points, we can see that the trend is predominantly linear.

Now we have discussed how the data is linear with respect to all variables, we may wish to fit linear models to the data. Since there are only two predictor variables, to perform an exhaustive check all models, we require three different linear models,

$$\begin{aligned}
\text{Length} &= \beta_0 + \beta_1\text{Height} + \beta_2\text{Weight} \\
\text{Length} &= \beta_0 + \beta_1\text{Height} \\
\text{Length} &= \beta_0 + \beta_1\text{Weight}.
\end{aligned}$$

These linear models will be fit using the lm() function built into R as follow,

```
lm1<-lm(Length~Height+Weight, data=catheter)
lm2<-lm(Length~Height, data=catheter)
lm3<-lm(Length~Weight, data=catheter)
```

For linear models of this form there are four assumptions to check, for each model, these assumptions are: linearity, homoscedasticity, normality and independence. It should be noted for multiple linear regression, we need to check for linearity and homoscedasticity between the residuals and the overall fitted data as well as between residuals and each predictor variable. There is no need to do this for simple linear regression as the residuals vs fitted plot is simply the residuals against the one predictor variable.

Now we will consider the assumptions for the first model, Length $= \beta_0 + \beta_1$Height $+ \beta_2$Weight. First, we must create the plots required for checking these assumptions hold.

```
res1<-residuals(lm1)
par(mfrow=c(3,2))
plot(lm1,which=1)
plot(lm1,which=2)
plot(lm1,which=3)
plot(lm1,which=5)
plot(catheter$Height, res1,xlab = "Height",ylab="Residuals",main="Residuals vs Height")
plot(catheter$Weight, res1,xlab = "Weight",ylab="Residuals",main="Residuals vs Weight")
```

Linearity

To test for linearity between residuals and fitted values, consider the top left hand plot in Figure 2, titled residuals vs fitted. If the data is linear, we would expect an equal number of points above and below the line '$y = 0$' for each 'x' value, with no curvature. For this data set, we can see that the data follows a reasonably linear trend with an equal spread of values above and below the horizontal line. Normally, the included red line may be helpful when discussing linearity, in this plot, the line clearly isnt linear, however this is due to the sheer lack of data. Overal there is insufficient evidence in the plot to invalidate the assumption of linearity.

To check for linearity between the residuals and the individual predictor variables, we need to consider the two bottom plots in Figure 2. First consider residuals vs Height, where we can see the data follows a similar trend to that described in the residuals vs fitted plot, once again due to the lack of data points, the assumption of linearity holds. Second the residuals vs Weight plot, we can see that the data follows a similar trend to

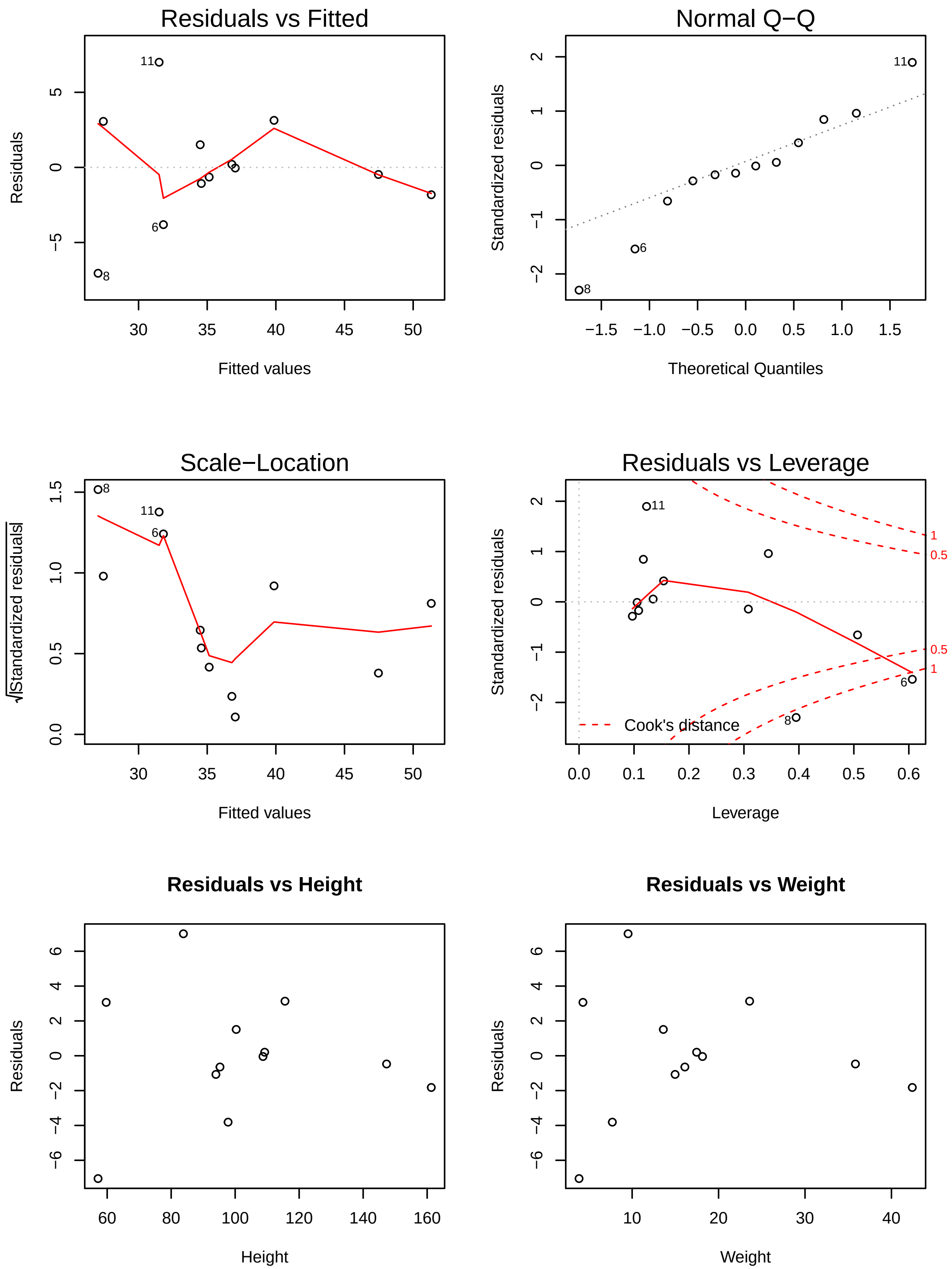

Figure 2: The six plots required to thoroughly check the assumptions for the first linear model.

the other two plots considered, thus the same justification can be applied to claim that the assumptions of linearity holds for the whole model.

Homoscedasticity

To test for homoscedasticity between residuals and fitted values, consider the top left hand plot in Figure 2, titled residuals vs fitted. If the data is homoscedastic, we would expect to see an equal distance between the points and the horizontal line described above. However for this data set, we have that the plot has significant fanning, where the variance decreases as the the Heigth and Weight increase. This is not all that surprising since the data is of predominantly young children, with only two data points corresponding to large catheters. This explains why the variance is smaller for these values. Given the small number of data points, the assumption of homoscedasticity holds, albeit weakly. Another way to check for constant variance is to consider the scale location (center left) plot and look for linearity in the data points. However for this data set, we can see some curvature, which corresponds to the variance not being constant.

Now, to check homoscedasticity between residuals and height and weight independently, we need to consider the two bottom plots of 2. In both of these plots, we can see that the data follows a similar trend to that of the plot discussed above, with significant fanning. However this fanning will be due to the small data set, so once again the assumption of constant variance holds.

Normality

To test for Normality, consider the Normal-QQ (top right) plot in Figure 2. If the residuals are Normally distributed, we would expect to see a straight, linear line following the tend of $y = x$. For this data set, we can see that in general it follows a straight line, with some significant deviation from the trend for lower values. This significant deviation implies some negative skewing in the distribution. However since there are only 12 data points, randomness that is intrinsic in this plot is not unexpected and is unavoidable. Thus given the small data set, the residuals appear to be passably normal, thus there is insufficient evidence to invalidate this assumption.

Independence

There is no formal test for independence in the data, instead we need to consider the way the data was collected. This data set was collected from children with congenital heart defects, implying there is a slim chance that there is any relationship between the individuals. Thus it appears that there is no connection between each subject, thus the data is independent to the best of our knowledge.

Now we have checked the assumptions for the multiple regression model, we also need to check the same assumptions for the two simple regression models. However since we have discussed them in considerable detail above, we will breifly discuss them for the simple models.

For the simple regression model in just Height, we have the following diagnostic plots.

```r
par(mfrow=c(2,2))
plot(lm2)
```

Using Figure 3, we will check the following assumptions,

- *Linearity*
  - In the residuals vs fitted plot, the data follows a reasonably linear trend with an almost even spread of points above and below the horizontal line. Although this lack of linearity can be explained by the small data set, hence the data is approximately linear and the assumptions holds.
- *Homoscedasticity*
  - In the residuals vs fitted plot, there is considerable fanning in the points, particularly for the smaller fitted values. Similarly in the scale location plot the data has some curvature present in it further implying a non-constant variance. However due to the small data set, this is insufficient evidence to invalidate the assumption of homoscedasticity.
- *Normality*

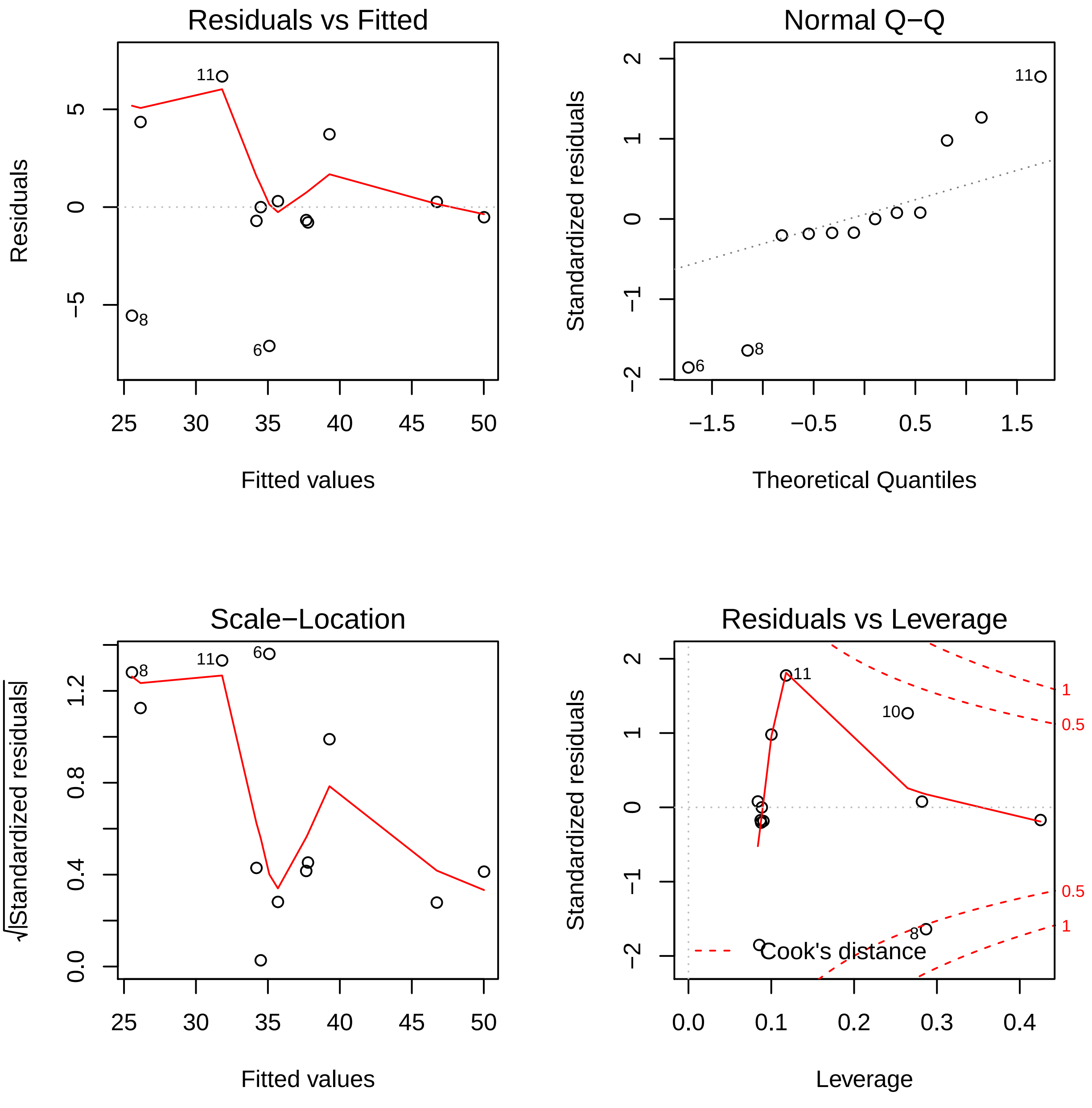

Figure 3: The four plots required to check the assumptions for the second linear model.

  - In the normal-QQ plot, we can see that the data follows an almost linear trend with significant deviation at the ends. However the small data set has an even bigger impact on normality as it does not satisfy the conditions of the central limit theorem, so we have insufficient evidence to invalidate the assumptiosn of normality.

- *Independence*
  - The data used to create this model is the same as that used to create the multiple regression model. Since the assumption of independence is satisfied in the previous model it is also satisfied in this model.

It should be noted that there is one point in this model that has significant leverage, however once again due to the small number of data points, one outlier has a much bigger affect than for larger data sets.

For the simple regression model in just Weight, we have the following diagnostic plots.

```
par(mfrow=c(2,2))
plot(lm3)
```

Using Figure 4, we will check the following assumptions,

- *Linearity*
  - In the residuals vs fitted plot, the data follows a reasonably linear trend with an almost even spread of points above and below the horizontal line. Although this lack of linearity can be explained by the small data set, hence the data is approximately linear and the assumptions holds.
- *Homoscedasticity*
  - In the residuals vs fitted plot, there is considerable fanning in the points, particularly for the smaller fitted values. Similarly in the scale location plot the data has some curvature present in it further implying a non-constant variance. However due to the small data set, this is insufficient evidence to invalidate the assumption of homoscedasticity.
- *Normality*
  - In the normal-QQ plot, we can see that the data follows an almost linear trend with some deviation at the ends. However the small data set has an even bigger impact on normality as it does not satisfy the conditions of the central limit theorem. However this plot follows a significantly stronger trend of normality than the previous model discussed.
- *Independence*
  - The data used to create this model is the same as that used to create the multiple regression model. Since the assumption of independence is satisfied in the previous model it is also satisfied in this model.

It should be noted that there is one point in this model that has significant leverage, however once again due to the small number of data points, one outlier has a much bigger affect than for larger data sets. Interestingly this one point in both models corresponds to the same child, a 3.86kg baby.

After checking the assumptions hold for each linear model we wish to compare the coefficients of the predictor variables in the different linear models. Below are summaries of the three linear models as above.

```
summary(lm1)
```

```
##
## Call:
## lm(formula = Length ~ Height + Weight, data = catheter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

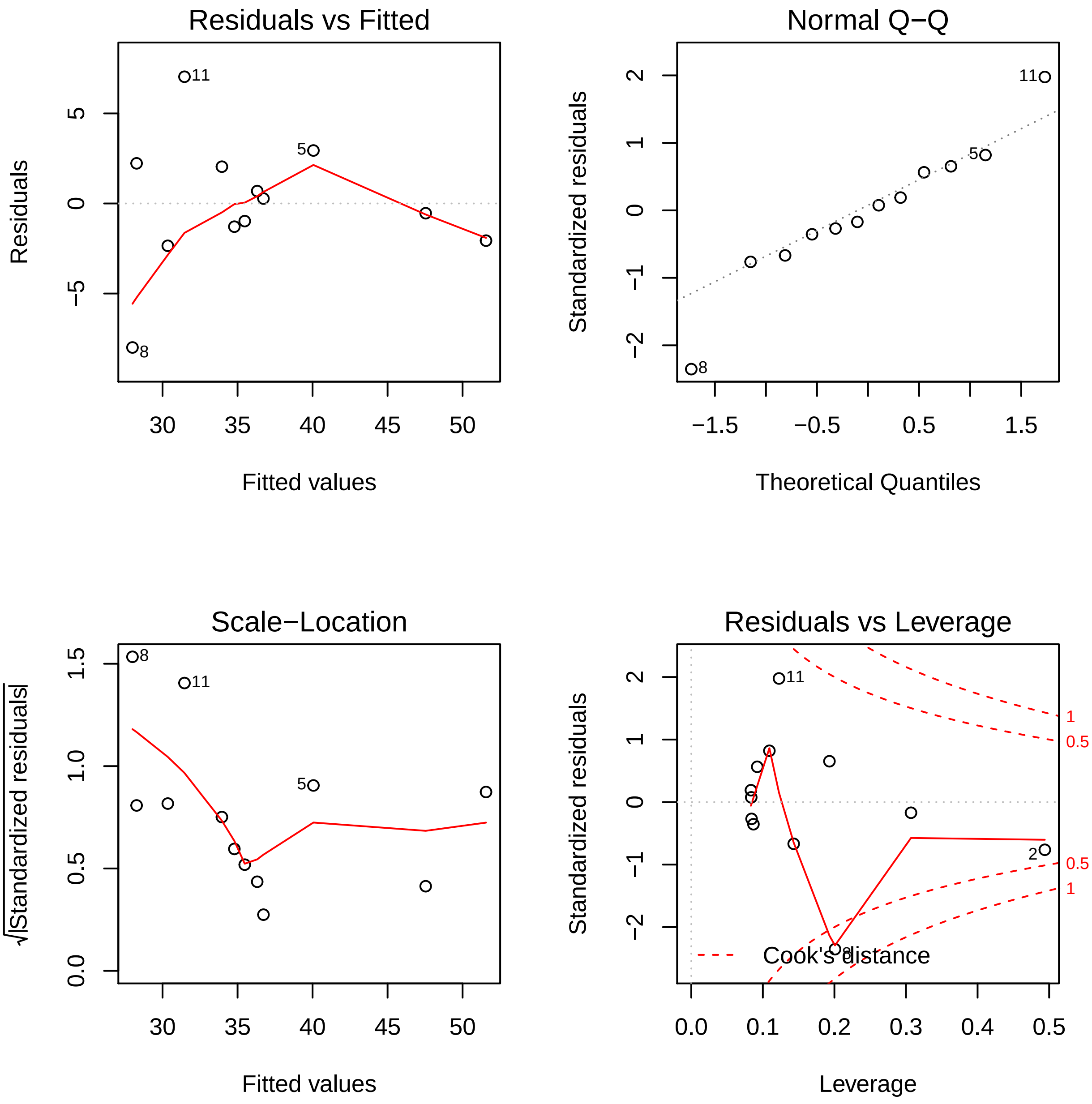

Figure 4: The four plots required to check the assumptions for the third linear model.

```
## -7.0497 -1.2588 -0.2576  1.8987  7.0030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.00828    8.74782   2.402   0.0398 *
## Height       0.07729    0.14192   0.545   0.5993
## Weight       0.42081    0.36405   1.156   0.2775
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.943 on 9 degrees of freedom
## Multiple R-squared:  0.8054, Adjusted R-squared:  0.7621
## F-statistic: 18.62 on 2 and 9 DF,  p-value: 0.0006332
```

```r
summary(lm2)
```

```
##
## Call:
## lm(formula = Length ~ Height, data = catheter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0996 -0.7246 -0.2608  1.1585  6.6826
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.12402    4.24711   2.855 0.017113 *
## Height       0.23495    0.03986   5.894 0.000152 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.008 on 10 degrees of freedom
## Multiple R-squared:  0.7765, Adjusted R-squared:  0.7541
## F-statistic: 34.74 on 1 and 10 DF,  p-value: 0.0001523
```

```r
summary(lm3)
```

```
##
## Call:
## lm(formula = Length ~ Weight, data = catheter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9958 -1.4818 -0.1334  2.0899  7.0378
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.63596    2.00425  12.791 1.60e-07 ***
## Weight       0.61136    0.09698   6.304 8.86e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.801 on 10 degrees of freedom
## Multiple R-squared:  0.7989, Adjusted R-squared:  0.7788
## F-statistic: 39.74 on 1 and 10 DF,  p-value: 8.865e-05
```

First we will consider the coefficients of Height, in the simple linear regression, we have $\beta_1 = 0.235$, where as in the multiple linear regression, we have, $\beta_1 = 0.077$. The significant difference in these values can be explained by the inclusion of Weight in the linear model. For the simple linear regression model, the intercept coefficient is, $\beta_0 = 12.124$, where as in the multiple linear regression, the intercept coefficient is, $\beta_0 = 21.008$. If we niavely ignore the inclusion of Weight in the model, the larger coefficient of Height in the simple linear regression corresponds to a smaller intercept and vice versa.

When we include Weight in the multiple regression model, it has coefficient, $\beta_2 = 0.421$, in constrast to the simple regression model where the coefficient is, $\beta_2 = 0.611$. Once again the large coefficient corresponds to a smaller intercept coefficient and vice versa.

Now for both simple regression models, the coefficients of Height and Weight (independently) are statistically significant and should be included in the model. However for the multiple regression model, both coefficients are statistically insignificant implying a strong correlation between these two predictors. If we were performing model selection, we would choose to drop one of the non-significant predictor variables in the multiple regression model and simplify it to one of the single regression models described above.

For the simple regression model containting Weight, we can interpreted the coefficient to mean that for any child, an increase in Weight by one (in kilograms) will result in an increase in length of the catheter tube by 0.611cm.

Similalry, for the multiple regression model the coefficient of Weight can be interpreted as, for an increase in one (kilogram) of Weight for a child of fixed Height, the length of the catheter tube will increase by 0.421cm.

Up until now, we have considered the multiple linear regression as a linear function of the data, however it can also be thought of as projecting the data onto model subspaces. Let $\mathcal{L}_1$ be the model subspace corresponding to the simple linear regression with Height and $\mathcal{L}_2$ be the model subspace corresponding to the simple lienar regression with Weight.

It is known that $\mathcal{L}_1$ is the column space of the model matrix corresponding to the simple regression model it describes. Thus,

$$\mathcal{L}_1 = \text{span}\left\{\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}, \begin{pmatrix}108.7\\161.29\\\vdots\\147.32\end{pmatrix}\right\}.$$

Similarly, $\mathcal{L}_2$ is the column space of the model matrix corresponding to the simple regression model it describes. Hence we have,

$$\mathcal{L}_2 = \text{span}\left\{\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}, \begin{pmatrix}18.14\\42.41\\\vdots\\35.83\end{pmatrix}\right\}.$$

Now consider the intersection of these two subspaces,

$$\begin{aligned}\mathcal{L}_1 \cap \mathcal{L}_2 &= \left\{\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}, \begin{pmatrix}108.7\\161.29\\\vdots\\147.32\end{pmatrix}\right\} \cap \left\{\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}, \begin{pmatrix}18.14\\42.41\\\vdots\\35.83\end{pmatrix}\right\}\\ &= \text{span}\left\{\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}\right\}\\ &= \text{span}\{\mathbf{1}\}\end{aligned}$$

Thus,

$$\{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp} \quad = \quad \{\mathbf{1}\}^{\perp}.$$

Now we need to consider the following subspaces,

$$\begin{aligned}
\mathcal{L}_1 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp} &= \text{span}\left\{\begin{pmatrix} 108.7 \\ 161.29 \\ \vdots \\ 147.32 \end{pmatrix}\right\} \\
&= \boldsymbol{h} \\
\mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp} &= \text{span}\left\{\begin{pmatrix} 18.14 \\ 42.41 \\ \vdots \\ 35.83 \end{pmatrix}\right\} \\
&= \boldsymbol{w}
\end{aligned}$$

These subspaces are the components in the models that are orthogonal to the intersection of the two models. That is to say it is the components that are not shared by either model, so if these two subspaces are orthogonal then the two models are disjoint.

Since these two subspaces are one-dimensional, we can calculate the angle between these two subspaces as follows.

$$\begin{aligned}
\theta &= \cos^{-1}\left(\frac{\boldsymbol{h}^T\boldsymbol{w}}{||\boldsymbol{h}||||\boldsymbol{w}||}\right) \\
&= 0.322 \quad \text{radians.}
\end{aligned}$$

Since the angle between these two subspaces is not $\frac{\pi}{2} \approx 1.57$, i.e they are not perpendicular, the two predictor variables are correlated. This reuslt is not surprising since taller children are generally heaver and vice versa.

Now we have checked the model assumptions and discussed the angle between the subspaces of the simple models. We can now select a final model to be used in predicting the lengths of catheter used. In the multiple regression model fitted above, we saw that the coeffficients were statistically insignificant. This was then explained by looking into the angle between the two vector subspaces for Heigth and Weight, which showed that Height and Weight were moderately correlated. So it seems unnecessary to fit both Height and Weight as predictors, so to prevent overfitting we will look at using only the two predictor variables.

As far as distinguishing between the two simple regression models, the AIC statistic was considered, as AIC weighs up goodness of fit and the number of predictors, however since the number of predictors is the same between the two models, AIC is just a measure of goodness of fit. The two simple regression models, in terms of Height and Weight have AIC values, 71.19 and 69.92 respectively. Although the difference in these two values is very small, it implies that the simplre regression model as a function of Weight is slightly better than Height.

Similarly, we can look at plots of the data set with the prediction intervals overlaid, for both models to see if there is an obvious difference in the width or shape of these intervals. See figure 5.

```
par(mfrow=c(1,2))

predicts.lm2 <- predict(lm2,interval="prediction")
plot(catheter$Height,predicts.lm2[,1],xlab = 'Height',ylab='predicted value',main='Predicted value agai
lines(lowess(catheter$Height,predicts.lm2[,2]))
lines(lowess(catheter$Height,predicts.lm2[,3]))

predicts.lm3 <- predict(lm3,interval="prediction")
plot(catheter$Height,predicts.lm3[,1],xlab = 'Weight',ylab='predicted value',main='Predicted value agai
lines(lowess(catheter$Height,predicts.lm3[,2]))
lines(lowess(catheter$Height,predicts.lm3[,3]))
```

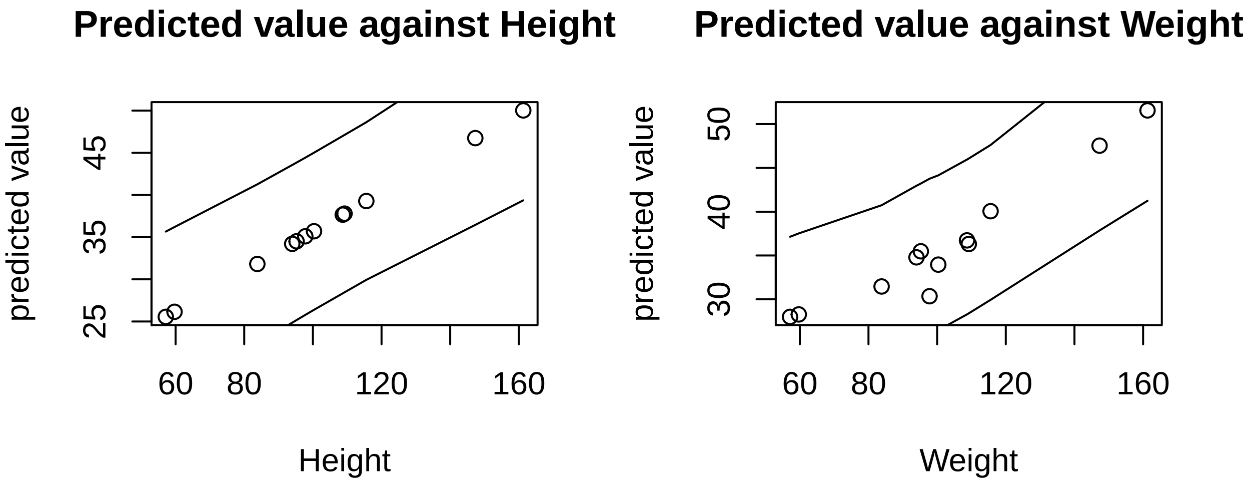

Figure 5: Prediction intervals plotted over the data for the two models under consideration.

Here we can see that the two plots are almost identical with a slightly tighter prediction interval for the model in terms of Weight. However as a result, the prediction interval is slightly wider for more extreme values of Weight.

Another approach considered in determining which model we should choose is cross validation, however this approach requires a partitioning the data into training and testing sets. Since we have only 12 data points, this partitioning will leave us with too little training data, so this idea was scrapped.

All that has been discussed above highlights that the two simple regression models are nearly identical. However the model corresponding to Height, satisfied the assumptions for linear regression slightly better than the model predicted by Weight. Thus the model we will choose to make the final predictions will be the second model, Length ∼ Height.