ST512 – Lab 02 *Solutions*

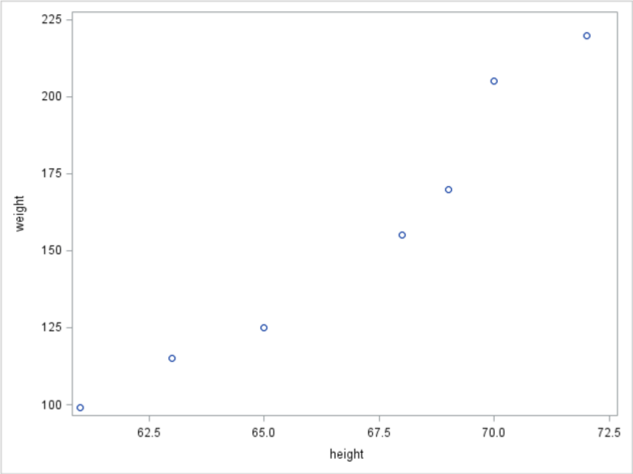
Part 1: Simple Linear Regression

See the solutions to this part on Lab 01.

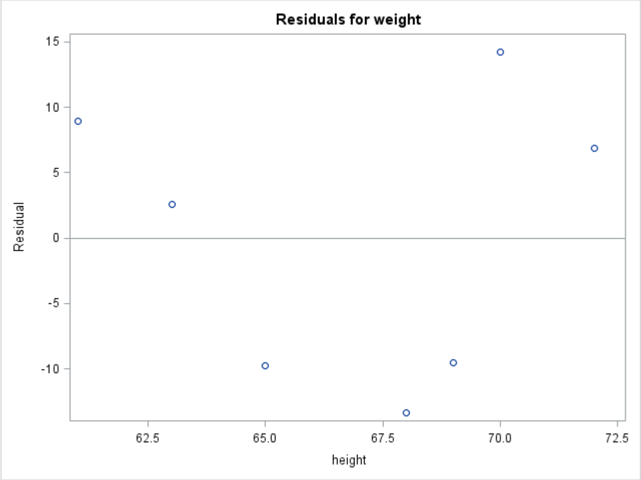
Part 2: Beyond the Linear Model

1. Draw a scatterplot of `weight` versus `height`. Is there a linear relationship?

   A scatterplot of `weight` versus `height` can be seen in the figure below. A linear relationship between the two variables isn't out of the question, but there are some mild signs of curvature.



2. Fit a linear regression model. Draw a plot of the residuals versus `height`. Do you see that there is a subtle quadratic-like pattern?

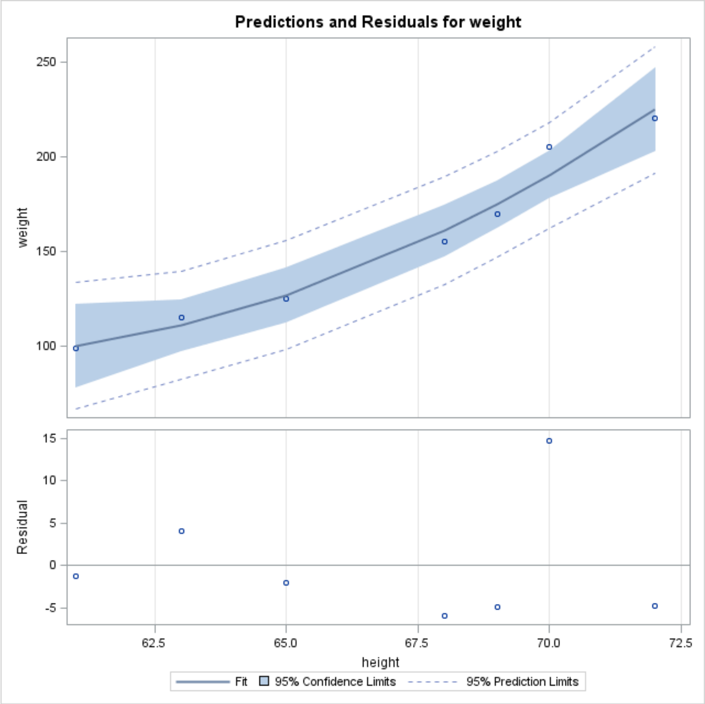


   From the residual plot above, it is more apparent that a linear relationship does not best describe the data. The residuals show a quadratic-like pattern when plotted versus `height` with positive residuals at either extreme and negative residuals clustered at the center of the data.

3. Re-draw the scatterplot of `weight` versus `height` and overlay the estimated quadratic regression curve. Comment on the fit. Also plot the residuals of this new fit versus `height` and notice that the pattern observed before is gone.

   The scatterplot of `weight` versus `height` with the overlayed regression curve and the residual plot for the quadratic regression are shown in the figure below. The quadratic regression curve appears to fit the data very well. In addition to seeing this in the scatterplot, we can also look at R^2. For the linear model, $R^2 = 0.9441$ and for the quadratic model $R^2 = 0.9743$. The higher R^2 value for the quadratic regression model further illustrates that this model fits the data very well.

   When looking at the residual plot for the quadratic regression model, there is no longer a noticeable quadratic-like pattern as was the case with the residuals from the linear regression model. Instead, the residual plot of the quadratic regression model does not appear to have a noticeable pattern to it.



PART 3: TRANSFORMATIONS

1. Draw a scatterplot of `rate` versus `dose`. Is there a linear relationship?

   The scatterplot of `rate` versus `dose` is shown on the left in Figure 1. It doesn't seem like there is a linear relationship. The slope appears to be greater for smaller doses and then gets smaller with higher doses.

2. Define a new variable `log.dose` as the natural log of `dose` and draw a scatterplot of `rate` versus `log.dose`. Does it look linear?

   The scatterplot of `rate` versus `log.dose` is shown on the right in Figure 1. It appears

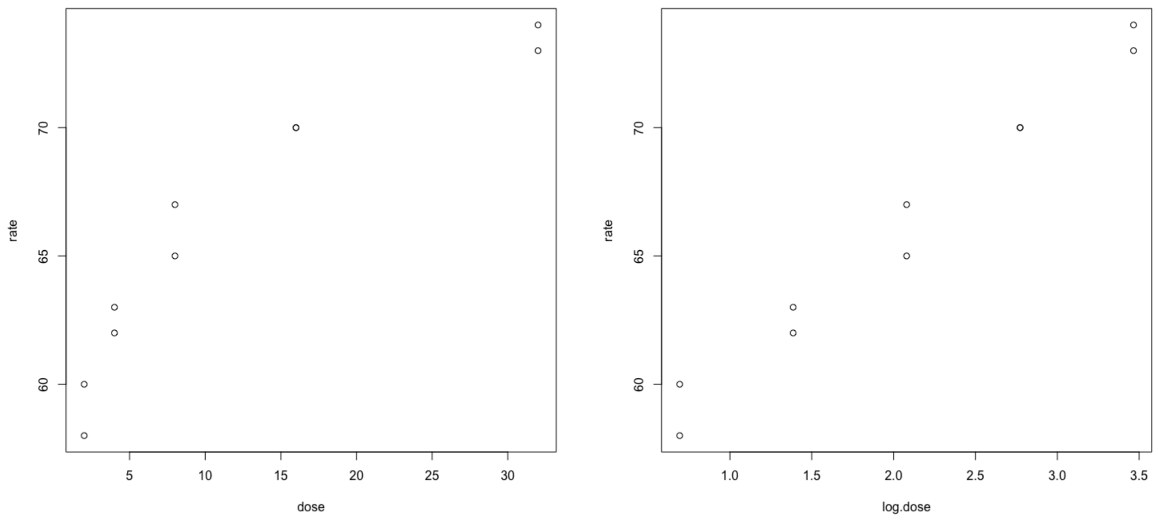


Figure 1: Scatter plots of `rate` versus `dose` (left) and `rate` versus `log.dose` (right).

to be a linear relationship between `rate` and `log.dose`.

3. Fit a linear model of `rate` versus `log.dose`. Comment on the quality of the fit. Interpret the estimated coefficient on `log.dose` in terms of the original variables.

   The R output is below.

```
Coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  55.2500     0.5950   92.85 2.02e-13 ***
log.dose      5.2658     0.2588   20.34 3.56e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.8023 on 8 degrees of freedom
Multiple R-squared:  0.981,Adjusted R-squared:  0.9787
F-statistic: 413.9 on 1 and 8 DF,  p-value: 3.562e-08
```

   The R^2 of the linear model (`rate` versus `log.dose`) is 0.9810. This is very high, suggesting that the model fits the data very well. The estimated coefficient of `log.dose` from the model is 5.266.

   Interpretation of the coefficient is a bit different here than in the linear model, and here is an explanation of how that argument goes. Consider the estimated value of `rate` corresponding to a `dose`, d, and consider the estimated value of `rate` after doubling the `dose`. Using the linear regression model, we can write

$$y_1 = \hat{\beta}_0 + \hat{\beta}_1 \log(d) \quad y_2 = \hat{\beta}_0 + \hat{\beta}_1 \log(2d)$$

Subtracting these two equations gives the change in `rate` after doubling the `dose`:

$$y_2 - y_1 = \hat{\beta}_1\{\log(2d) - \log(d)\} = \hat{\beta}_1 \log(2)$$

Therefore, the interpretation of the estimated coefficient on `log_dose` is that the increase in `rate` associated with a doubling of `dose` is estimated to be $\hat{\beta}_1 \log(2) = 5.26585 \log(2) = 3.65$. You could also interpret the estimated coefficient based on other changes in `dose` as appropriate. For example, the estimated change in `rate` associated with a 10% increase in `dose` is $\hat{\beta}_1 \log(1.1)$.