MSAN 601 - Homework 2

Andre Guimaraes Duarte

September 12, 2016

1

See `HW2.R` for relevant code.

In this question, we regress Length of Stay (Y) on Average Daily Census (X) from `SENIC_data.csv`. Four linear regression models are constructed:

- $Y \sim X$,
- $192Y \sim 192X$,
- $47Y \sim X$,
- $Y \sim 12X$.

Table 1 shows the values obtained forb_0, b_1, and R^2 for each model.

Model	$Y \sim X$	$192Y \sim 192X$	$47Y \sim X$	$Y \sim 12X$
b_0	8.520 93	1636.018 74	400.483 76	8.520 93
b_1	0.005 89	0.005 89	0.276 88	0.000 49
R^2	0.224 57	0.224 57	0.224 57	0.224 57

Table 1: Summary of results for Question 1

We can see that the R^2 is the same regardless of the scalar multiplication made to X and/or Y. When X and Y are multiplied by the same scalar, the slope remains the same, but the intersect changes (it is multiplied by the same scalar). If only Y is multiplied by a scalar, both the slope and the intercept change in accordance to this multiplier. If only X changes, the intercept remains the same while the slope varies with the inverse of the scalar.

In all cases, the coefficient of determination remains unchanged: the sample variation in Y that is explained by the variation in X is constant. In addition, the test statistics (F^*, $t^*\{b_0\}$, and $t^*\{b_1\}$) all remain constant for all four models. This can be easily proven using the definitions of b_0, b_1, $SSTO$, SSE, and SSR (not included here, but appended at the end of the homework if relevant).

2

1 We use `platsicHardness.txt` to create a linear regression model of plastic hardness in Brinell units (Y) as a function of the number of hours elapsed since the plastic was molded (X). File `HW2.R` contains the relevant code for this question. A quick plot of the data shows a seemingly positive linear relation between X and Y.

2 The F-statistic and its associated p-value show that the linear model seems adequate for this study.

We get a coefficient of determination $R^2 = 0.9731$, meaning that 97.31% of the variance in plastic hardness is explained by the model.

We get the parameter estimates for the intercept and the slope: $b_0 = 168.6$ BHN and $b_1 = 2.03438$ BHN/h. They are both significantly different from 0, as seen from the associated p-values. This means that at time 0, the initial plastic hardness is equal to 168.6 Brinell units. After an hour has elapsed, the mean expected increase in plastic hardness is eauql to 2.03438 Brinell units.

In order to test the adequacy of the model, we investigate the residuals. The residual plot shows that the residuals seem homoskedastic. The sequential plot does not show any dependence between the residuals. In addition, no outliers seem to be present, as per the standardized and studentized residual plots. Finally, the residuals are normally distributed, as shown by a QQ-plot or a normality test such as Shapiro-Wilk's.

In conclusion, nothing tells us that the linear regression model is not appropriate for this study.

3 We use the `leveneTest` from the package `car` with the flag `center = median` in order to perform a Brown-Forsythe test to determine whether or not the error variances varies with the level of X. First, however, we divide the data into two groups, $X \leq 24$ and $X > 24$. Our α is equal to 0.05.

We obtain a p-value of $0.2402 > \alpha$. The test is not significant, so we have no reason to reject the null hypothesis that the residual variances are homogenous. This is in accordance to our previous checks.

3

Read.

4

1 We use `muscleMass.txt` to create a linear regression model of muscle mass (Y) as a function of age in women (X). File `HW2.R` contains the relevant code for this question. An initial plot of the data seems to show a decreasing relation between X and Y.

The F-statistic and its associated p-value show that the linear model seems adequate for this study.

We get a coefficient of determination $R^2 = 0.7501$, meaning that 75.01% of the variance in plastic hardness is explained by the model.

We get the parameter estimates for the intercept and the slope: $b_0 = 156.3466$ lbs and $b_1 = -1.19$ lbs/year. They are both significantly different from 0, as seen from the associated p-values. The intercept in this case does not have a lot of meaning (muscle mass at age 0) and is not important for the linear model anyway. The information that this model brings is that we expect to see a mean decrease of 1.19 lbs in muscle mass for every year in women.

In order to test the adequacy of the model, we investigate the residuals. The residual plot shows that the residuals seem homoskedastic, except for maybe one observation, number 53. The sequential plot does not show any dependence between the residuals. The standardized and studentized residual plots again show that there may be a single outlier in the data, but it is not clear. Finally, the residuals are normally distributed, as shown by a QQ-plot or a normality test such as Shapiro-Wilk's.

In conclusion, nothing tells us that the linear regression model is not appropriate for this study. There is a single point that could eventually be considered an outlier, but no immediate action is required.

2 We run a Breush-Pagan test to determine whether or not the error variances vary with the level of X, using $\alpha = 0.01$. From the package `lmtest`, we use the function`bptest` on our linear model. We obtain a p-value of $0.0348 < \alpha$. Therefore, we reject the null hypothesis that the residual variances are homogenous. This means that the residuals fail the homoskedasticity test, and the linear model we constructed should be reevaluated before continuing.

Note: `bptest` uses studentized residuals instead of actual residuals. If we use the function `ncvTest` from the package `car`, we obtain a p-value of $0.05073 \approx \alpha$. This result if very close to α, and it would be risky to draw a conclusion based solely on this test. This is a good example to show that it is always better to run several tests on the data instead of coming to hasty conclusions.

If we remove observation 53 (which would only be done after more serious consideration) and re-run the model, all the tests for the residuals pass (the p-value for the Breush-Pagan test is $0.1191 > \alpha$).

5

For SLR, we know that the confidence interval (CI) and the prediction interval (PI) for a new observation $\hat{Y}_h$ are given by

CI: $\hat{Y}_h \pm t_{(1-\frac{\alpha}{2};n-2)} s\{\hat{Y}_h\}$

PI: $\hat{Y}_h \pm t_{(1-\frac{\alpha}{2};n-2)} s\{pred\}$

where

$s^2\{\hat{Y}_h\} = MSE[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n X_i - \bar{X}}]$

$s^2\{pred\} = MSE + s^2\{\hat{Y}_h\} = MSE[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n X_i - \bar{X}}]$

Therefore, the difference in the upper limits is given by

$$\begin{aligned}
\hat{Y}_h + t_{(1-\frac{\alpha}{2};n-2)} s\{pred\} - (\hat{Y}_h + t_{(1-\frac{\alpha}{2};n-2)} s\{pred\}) &= t_{(1-\frac{\alpha}{2};n-2)} s\{pred\} - t_{(1-\frac{\alpha}{2};n-2)} s\{\hat{Y}_h\} \\
&= t_{(1-\frac{\alpha}{2};n-2)} [s\{pred\} - s\{\hat{Y}_h\}] \\
&= t_{(1-\frac{\alpha}{2};n-2)} \sqrt{MSE}[\sqrt{\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n X_i - \bar{X}}} - \sqrt{1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n X_i - \bar{X}}}]
\end{aligned}$$

In this case, we are using `SENIC_data.csv` and regressing Length of Stay on Infection Risk. We have

$\alpha = 0.03$,

$n = 113$,

$X_h = 11.93$,

$\hat{Y}_h = 15.40861$,

$t_{(1-\frac{\alpha}{2};n-2)} = 2.19835$,

$MSE = 2.63752$ ($\sqrt{MSE} = 1.62404$),

$s\{pred\} = 1.29380$,

$s\{\hat{Y}_h\} = 0.293797$.

We obtain the difference between the upper limit of the confidence interval and the upper limit of the prediction interval:

$\hat{Y}_h + t_{(1-\frac{\alpha}{2};n-2)} s\{pred\} - (\hat{Y}_h + t_{(1-\frac{\alpha}{2};n-2)} s\{pred\}) = 2.12578.$

6

1 See `linearBC.R`.

2 See `bisectionBC.R`.

Mathematical proofs for Question 1

We have
$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})\cdot(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$ and $b_0 = \bar{Y} - b_1 \cdot \bar{X}$.

If we multiply X by a constant c_1 and Y by a constant c_2 ($c_1 \neq 0$ and $c_2 \neq 0$), we get
$b_{1,model} = \frac{c_1 c_2 \sum_{i=1}^n (X_i - \bar{X})\cdot(Y_i - \bar{Y})}{c_1^2 \sum_{i=1}^n (X_i - \bar{X})^2}$ and $b_{0,model} = c_2\bar{Y} - c_1 b_{1,model} \cdot \bar{X}$.

- Therefore, we can see that for the model $cY \sim cX$ ($c_1 = c_2 = c$), we get
  $b_{1,model} = \frac{\sum_{i=1}^n (X_i - \bar{X})\cdot(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = b_1$ and $b_{0,model} = c(\bar{Y} - b_1 \cdot \bar{X}) = c \cdot b_0$.
  In this case, the slope of the new model is unchanged, and the intercept is multiplied by the constant.
- For the model $cY \sim X$ ($c_1 = 1$ and $c_2 = c$), we get
  $b_{1,model} = \frac{c\sum_{i=1}^n (X_i - \bar{X})\cdot(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = c \cdot b_1$ and $b_{0,model} = c\bar{Y} - b_{1,model} \cdot \bar{X} = c \cdot b_0$.
  In this case, the slope and the intercept are both multiplied by the constant.
- For the model $Y \sim cX$ ($c_1 = c$ and $c_2 = 1$), we get
  $b_{1,model} = \frac{\sum_{i=1}^n (X_i - \bar{X})\cdot(Y_i - \bar{Y})}{c\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{1}{c} \cdot b_1$ and $b_{0,model} = \bar{Y} - c \cdot b_{1,model} \cdot \bar{X} = b_0$.

In a similar fashion, it is easy to show the following results (the sum of squares deal with Y, so only c_2 comes into play):

$$\begin{array}{lclclcl} SSTO_{model} & = & c_2^2 \cdot SSTO & & & & \\ SSE_{model} & = & c_2^2 \cdot SSE & \Rightarrow & MSE_{model} & = & c_2^2 \cdot MSE \\ SSR_{model} & = & c_2^2 \cdot SSR & \Rightarrow & MSR_{model} & = & c_2^2 \cdot MSR \end{array}$$

Therefore, we get

- $R^2_{model} = \frac{SSR_{model}}{SSTO_{model}} = \frac{SSR}{SSTO} = R^2$,
- $F^*_{model} = \frac{MSR_{model}}{MSE_{model}} = \frac{MSR}{MSE} = F^*$,
- $s^2\{b_{1,model}\} = (\frac{c_2}{c_1})^2 s^2\{b_1\} \Rightarrow t^*\{b_{1,model}\} = \begin{cases} \frac{b_1}{\frac{c}{c}\cdot s\{b_1\}} = \frac{b_1}{s\{b_1\}}, & \text{if } c_1 = c_2 = c \\ \frac{c\cdot b_1}{c\cdot s\{b_1\}} = \frac{b_1}{s\{b_1\}}, & \text{if } c_1 = 1, c_2 = c \quad = t^*\{b_1\}, \\ \frac{b_1/c}{s\{b_1\}/c} = \frac{b_1}{s\{b_1\}}, & \text{if } c_1 = c, c_2 = 1 \end{cases}$
- $s^2\{b_{0,model}\} = MSE_{model}[\frac{1}{n} + \frac{c_1^2 \cdot \bar{X}^2}{c_1^2 \cdot \sum_{i=1}^n (X_i - \bar{X})^2}] = MSE_{model}[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}] = c_2^2 \cdot s^2\{b_0\}$.
  $\Rightarrow t^*\{b_{0,model}\} = \begin{cases} \frac{c\cdot b_0}{c\cdot s\{b_0\}} = \frac{b_0}{s\{b_0\}}, & \text{if } c_1 = c_2 = c \\ \frac{c\cdot b_0}{c\cdot s\{b_0\}} = \frac{b_0}{s\{b_0\}}, & \text{if } c_1 = 1, c_2 = c \quad = t^*\{b_0\}. \\ \frac{b_0}{s\{b_0\}}, & \text{if } c_1 = c, c_2 = 1 \end{cases}$