Homework #4 (Solutions)

STAT-UB.0003: Regression and Forecasting Models

Solutions adapted from N.S. Boudreau's *Instructor's Solution Manual* (2011).

1. MBS, Ex. 12.2. On part (f), do not find or interpret R_a^2.

Solution:

(a) $\hat{\beta}_0 = 506.346$; $\hat{\beta}_1 = -941.900$; $\hat{\beta}_2 = -429.060$

(b) $\hat{y} = 506.36 - 941.90x_1 - 429.1x_2$.

(c) SSE $= 151016$; MSE $= 8883$; $s = 94.251$.

We expect about 95% of the y values to be within $\pm 2s = \pm 188.502$ unites of the fitted regression equation.

(d) The p-value for $H_0 : \beta_1 = 0$ against the alternative $H_a : \beta_1 \neq 0$ is $p = .003$. Since $p < .05$, we would reject H_0; there is sufficient evidence to indicate $\beta_1 \neq 0$ at significance level $\alpha = .05$.

(e) The 95% confidence interval for β_2 is

$$\begin{aligned}\hat{\beta}_2 \pm t_{.025,n-k-1}\mathrm{SE}(\hat{\beta}_2) &= -429.060 \pm 2.110(379.83)\\ &= -429.060 \pm 801.4413\\ &= (-1230.5013, 372.3813).\end{aligned}$$

We have used that $n - k - 1 = 20 - 2 - 1 = 17$, so that $t_{.025,n-k-1} = 2.110$.

(f) $R^2 = 45.9\%$; the fitted regression model explains 45.9% of the variability in y.

(g) $F = 7.22$.

(h) The observed significance level is $p = 0.005$. Since the p-value is so small, we would reject $H_0 : \beta_1 = \beta_2 = 0$ for most values of the significance level α. We have very strong evidence that the model is useful (at least one of the predictor variables is useful for predicting y).

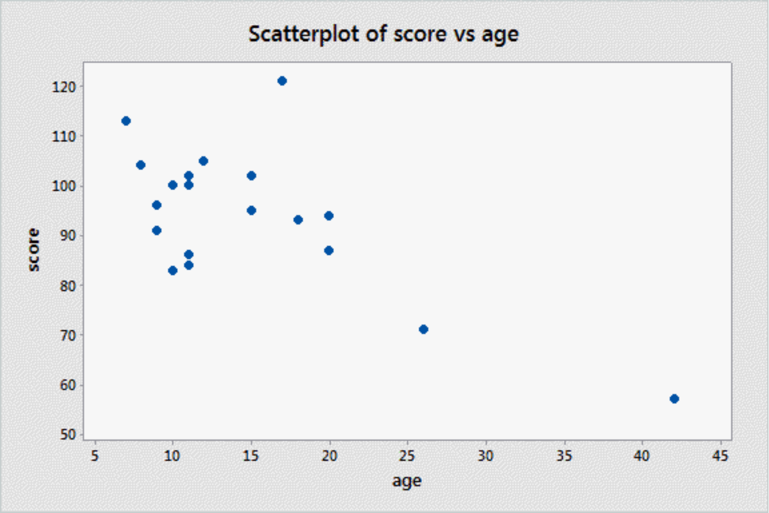
2. The `Gesell` dataset concerns a study of whether intelligence can be predicted based on the age at which a child starts to speak. For each of 21 participants in the study, the variable Age represents the age (in months) at which they spoke their first word, and the variable Score represents the Gesell Adaptive Score. (The Gesell test is an adult intelligence test).

   (a) Without looking at the data, how would you expect Score to be related to Age? (Positively or negatively?)

Solution: Negatively: children who learn to speak earlier likely have higher intelligence. (Full credit for any answer.)

(b) Make a scatterplot of Score versus Age. Does the plot show the relationship you predicted in (a)?

Solution: Here is the scatterplot:



This agrees with the relationship predicted in part (a).

(c) Run the simple regression of Score on Age. Get the leverage and Cook's Distance values by clicking on Storage in the regression dialog box, and checking the boxes for leverage (`Hi`) and Cook's Distance.

Solution: Here is the regression output:

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
11.0229  41.00%     37.89%      27.15%

Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant  109.87     5.07    21.68    0.000
age       -1.127    0.310    -3.63    0.002  1.00

Regression Equation

score = 109.87 - 1.127 age
```

(d) Use the regression output to compute the p-value for the coefficient of Age in the regression. Does this suggest that Score is related to Age? Does the sign of the fitted coefficient agree with your prediction.

Solution: $p = 0.002$. There is very strong evidence that score is related to age. The sign of $\hat{\beta}_1$ is negative, which agrees with our prediction.

(e) What proportion of the variability in Score is explained by Age, based on the regression output?

Solution: $R^2 = 41.00\%$

(f) Are there any data points with high leverage (`Hi` above $2k/n$)? Is the Cook's Distance corresponding to these points high enough (close to or above 1.0) to cause concern?

Solution: The point $(Age, Score) = (42, 57)$ has the highest leverage (.6516, which is above $2k/n = 2(1)/(21) = .09$). The Cook's distance for this point is .67811, below 1.0, which indicates the point does not have a strong impact on the regression.

(g) Delete the data point with the largest value of Cook's Distance, by highlighting that case in the Minitab worksheet, and pressing the Del key. Now, re-run the regression. Describe the effects on the p-value for the slope, and on R^2. Is there now strong evidence of a linear relationship between Score and Age?

Solution: Here is the updated regression model:

```
Model Summary

      S    R-sq  R-sq(adj)  R-sq(pred)
11.1068  11.22%      6.28%       0.00%

Coefficients

Term        Coef  SE Coef  T-Value  P-Value   VIF
Constant  105.63     7.16    14.75    0.000
age       -0.779    0.517    -1.51    0.149  1.00

Regression Equation

score = 105.63 - 0.779 age
```

The p-value increased to .149, and the R^2 decreased to 11.2%.

(h) Do you feel that it is justifiable to have deleted this point from the data set?

Solution:

No, this point should not have been deleted from the data set since it isn’t a bad leverage point and didn’t prove evidence of the usefulness of the regression model.âĂÍGiven the apparent regression line seems to depend on the outlier, we should consider a larger sample with larger predictor values if possible. (Full credit for any answer here.)