Statistics

Week 10 Recitation

ESD, SUTD

Term 5, 2017

Question 1 (Proof for R^2)

Show that in simple linear regression, the two formulas for r^2 agree, namely,

$$\frac{\mathsf{SSR}}{\mathsf{SST}} = \left(\frac{s_{xy}}{s_x s_y}\right)^2.$$

Question 1 (Proof for R^2 - Solution)

As $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_i$

We have $\hat{y}_i - \bar{y} = \hat{\beta}_1 (x_i - \bar{x})$.

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} \left(\hat{\beta}_1 (x_i - \bar{x})\right)^2 = \hat{\beta}_1^2 \, (n-1) s_x^2.$$

Also, $\text{SST} = (n-1) s_y^2$.

Therefore,

$$\frac{\text{SSR}}{\text{SST}} = \frac{\hat{\beta}_1^2 \, s_x^2}{s_y^2} = \frac{\left(s_{xy}/s_x^2\right)^2 s_x^2}{s_y^2} = \left(\frac{s_{xy}}{s_x s_y}\right)^2.$$

Question 2 (ANOVA)

Refer to the spreadsheet 'IQ', which tabulates some test subjects' IQ vs their brain size, height, and weight.

Is any of these predictors good at modeling IQ?

Hint

- Use the *Data Analysis* function in Excel
- Observe the *ANOVA* Table
- Null hypothesis of ANOVA: $\beta_1 = \beta_2 = ... = \beta_k = 0$

Answer:
P-value of F test is large (0.367). Thus do not reject H_0.
None of the three predictors are good at modeling IQ.

Question 3 (Polynomial Regression)

Refer to the spreadsheet 'weightlifting', which contains world records data in weightlifting for various weight categories. Construct polynomial regression models for the data.

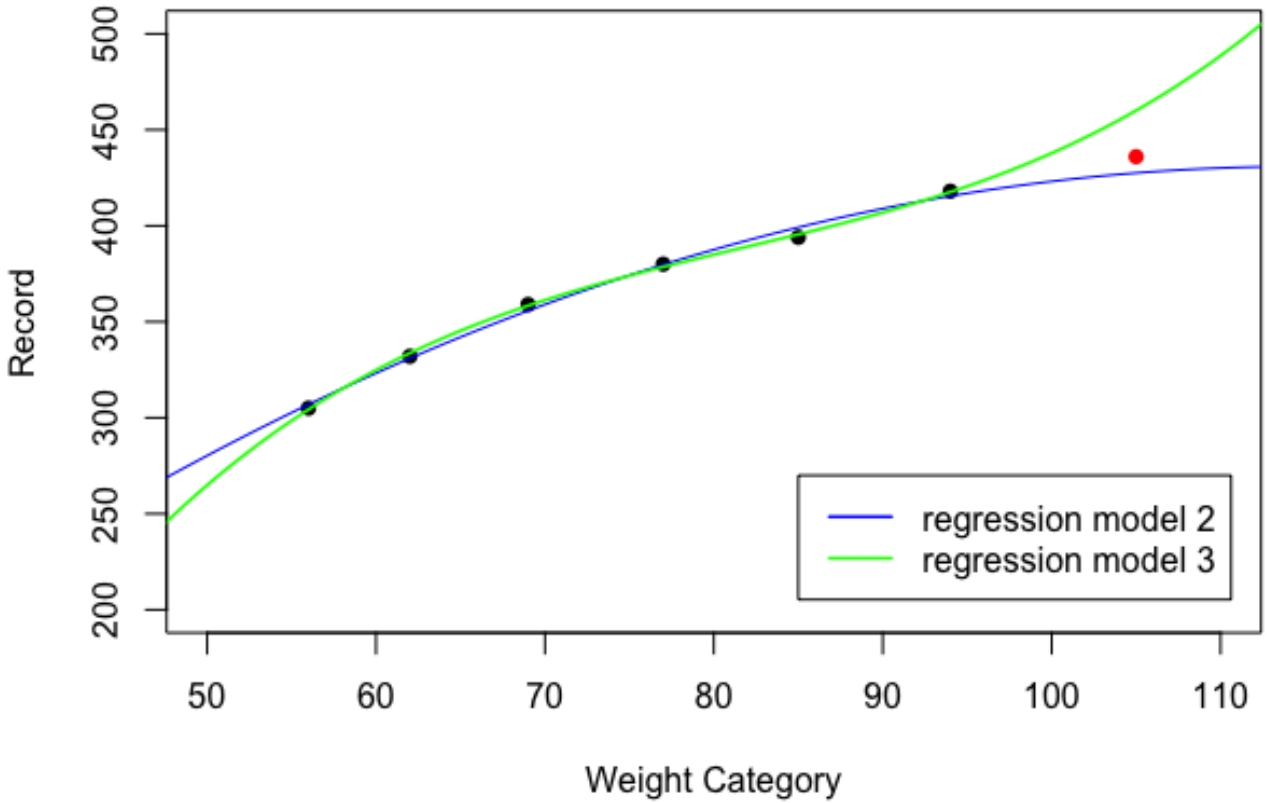
Compare these three models:

- $y = \beta_0 + \beta_1 x$
- $y = \beta_0 + \beta_1 x + \beta_2 x^2$
- $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

Which model is the best, in terms of the adjusted r^2 ?
Which model is the best, in terms of making a prediction ?

You can use the first six data points to fit the model, and use the last data point to validate the model.

Question 3 (Polynomial Regression - Solution)



Model 3 has smaller adjusted R^2.

However, model 2 makes more accurate prediction.

Question 4 (More Adjusted r^2)

Refer to the spreadsheet 'utilities', which contains data for average utilities consumption vs average property size for various types of properties in Singapore.

Our goal is to predict size based on one or more of the consumption columns. Find the best subset of predictors by maximizing adjusted r^2.

Answer:

Model $y_{size} = \beta_0 + \beta_1 x_{gas}$ has the highest adjusted r^2.

The correlation between electricity, water and gas are all extremely high. Therefore, to avoid multicollinearity and redundancy, it is wise to use only one of the predictors.