Notes on Linear Regression

Earl Patrick B. Macalam

December 11, 2020

1 Questions that Regression can Address

1. Is there a relationship between advertising budget and sales?
2. How strong is the relationship between advertising budget and sales?
3. Which media contribute to sales?
4. How accurately can we estimate the effect of each medium on sales? For every
5. How accurately can we predict future sales?
6. Is the relationship linear?
7. Is there synergy among the advertising media?

2 Assessing the Accuracy of the Model

1. Residual Standard Error (RSE)
   - an estimate of the standard deviation of ϵ. Roughly speaking, it is the average amount that the response will deviate from the true regression line
   - If the RSE is 3.26 then actual sales in each market deviate from the true regression line by approximately 3,260 units, on average.
   - If the predictions obtained using the model are very close to the true outcome values, that is, if $\hat{y_i} \approx y_i$ for $i = 1, ..., n$ then RSE will be small, and we can conclude that the model fits the data very well.
2. R^2
   - It takes the form of a proportion, the proportion of variance explained and so it always takes on a value between 0 and 1, and is independent of the scale of Y.
   - An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression.

- A number near 0 indicates that the regression did not explain much of the variability in the response; this might occur because the linear model is wrong, or the inherent error σ^2 is high, or both.
- An R^2 was 0.61 under two-thirds of the variability in sales is explained by a linear regression on TV.

3 Interpretation

3.1 Simple Linear Regression

- An additional $1,000 spent on TV advertising is associated with selling approximately 47.5 additional units of the product

3.2 Multiple Linear Regression

- We interpret these results as follows: for a given amount of TV and newspaper advertising, spending an additional $1,000 on radio advertising leads to an increase in sales by approximately 189 units.

3.3 On p-value of β

- We reject the null hypothesis, that is, we declare a relationship to exist between X and Y, if the p-value is small enough.
- Hence, if we see a small p-value, then we can infer that there is an association between the predictor and the response.

4 Diagnostics

1. The relationship between the IVs and the DV is linear.
   - Use residual plot
2. There is no multicollinearity in your data.
   - VIF, > 10 QUESTIONABLE
3. The values of the residuals are independent.
   - Durbin watson statistic and Breusch Godfrey Test
   - H_0: There is no correlation among the residuals.
     H_a: The residuals are autocorrelated
4. The variance of the residuals is constant.
   - Scale-Loc Plot

5. The values of the residuals are normally distributed.
    - Normal QQ plot
6. There are no influential cases biasing your model.
    - Cooks distance