STAT0028 Week 3 Exercises

Andrew Wu

2024-10-19

Question 1

(a)

We may refer to their predictors by x_i and responses by y for convenience.

- x_2 appears uncorrelated with x_3 and x_1, though three distinct groups emerge. Similar behaviour occurs with the response, thus we might infer that the data was taken from three different locations, where the average soil moisture tension is the same within each location.
- x_1, x_3 are positively correlated, and both are also positively correlated with the response
- There are no obvious outliers in the dataset

(b)

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e_i, \quad i = 1, \dots 27$$

(c)

Under a power of 5%, the p-value of the F-statistic in model 1 is statistically significant, so we reject the null hypothesis that none of the predictors provide useful information about the response.

(d)

Based on only the R-squared and the residual plots, model 1 would appear to have a decent fit. The adjusted R-squared of 0.7229 is not low, so the model offers decent predictive behaviour. The residuals split into two main groups based on average soil moisture tension and do are not in clear violation of homoskedasticity. Observations 15,18 are influential and we have some mild curvature at the tails possibly indicating non-normality, but this is not conclusive. The fitting of model 2 was likely prompted by the partition of the fitted values into two clear groups based on soil moisture tension. Squaring this predictor aligns the two lower values of moisture tension closer to 0.

(e)

The adjusted R-squared has improved to 0.8941 suggesting a much better fit. The residuals no longer separate so clearly into two groups when plotted against the fitted values and we have retained homoskedasticity. Normality of the residuals seems well satisfied except perhaps at the positive tail with two influential observations.

(f)

Based on a 5% significance level, the p-value of `I(x2^2)` is statistically significant and suggests we can reject the null hypothesis that the predictor gives no useful information about the response.

(g)

We can try dropping `x1` since it is positively correlated with `x3`, has a statistically insignificant p-value, and both OLS estimates have the same sign. The effects of this predictor should ideally be mostly captured by `x3`.

(h)

$$y_i = \beta_0 + \beta_1 x_2 + \beta_2 x_3 + \beta_3 x_2^2 + e_i$$

Models 2 and 3 clearly out perform Model 1 based on the fit (adjusted R-squared) and accounting for the two distinct groups based on soil moisture as well as the previously mentioned reasons in part (e). I would choose Model 3 over Model 2 for the following reasons:

- Maintains the explanatory power of model 2 (comparable adjusted R-squared of 0.89) whilst providing a simpler model by dropping the predictor x1, which was positively correlated with x3 and not statistically significant.
- Residuals still look homoskedastic on the residuals vs fitted values plots.
- The residuals look more normal on qq-plot, with no significant deviations at the tail, even for influential observations.

Question 2

Note, throughout this question, we make use of the fact that if A is invertible, then $(A^\top)^{-1} = (A^{-1})^\top$.

(a)

$$\begin{aligned} x_i^\top \beta_Y + e_i &= x_i^\top A^{-1}(a\beta_Z + b) + e_i \\ &= a(u_i^\top \beta_Z + d_i) + u_i^\top b \\ &= y_i. \end{aligned}$$

(b)

$$\begin{aligned} \hat{\beta}_y &= (X^\top X)^{-1} X^\top y \\ &= (A^\top U^\top U A)^{-1} A^\top U^\top (aZ + Ub) \\ &= A^{-1}(U^\top U)^{-1}(A^\top)^{-1} A^\top U^\top (aZ + Ub) \\ &= A^{-1}(a(U^\top U)^{-1} U^\top Z + b) \\ &= A^{-1}(a\hat{\beta}_Z + b) \end{aligned}$$

(c)

Observe that the hat matrices for both linear models are the same:

$$\begin{aligned} H &:= X(X^\top X)^{-1} X \\ &= UA(A^\top U^\top U A)^{-1} A^\top U^\top \\ &= U(U^\top U)^{-1} U^\top. \end{aligned}$$

It suffices to show that $\text{RSS}_Y = a^2 \text{RSS}_Z$.

$$
\begin{aligned}
\mathrm{RSS}_Y &= ((I-H)Y)^\top (I-H)Y \\
&= Y^\top (I-H)Y \quad I-H \text{ is symmetric and idempotent} \\
&= a^2 Z^\top (I-H)Z + 2ab^\top U^\top (I-H)Z + b^\top U^\top (I-H)Ub \\
&= a^2 Z^\top (I-H)Z \quad \text{as } U^\top (I-H) = 0 \\
&= a^2 \mathrm{RSS}_Z.
\end{aligned}
$$