DATA 303/473 Test 1: Graduate earnings - Questions

Nicholas Tran, 300296259

07/04/2022

1.

a.

`SAT` (median SAT score) and `price_with_aid` (price after financial aid) are excluded, They're most likely excluded due to `SAT` having really strong multicollinearity with `ACT` and `price_with_aid` having really strong multicollinearity with `price`.

b.

Residuals vs. Fitted: There's no major curvature to the points so there's no strong evidence of non-linearity.

Q-Q: There's a slight deviation at one of the tails, however not strong enough for any evidence of a departure from normality.

Scale-Location: All residuals appear to have equal spread so unequal variance assumption holds.

Residuals vs. Leverage: No evidence of outliers or influential observations.

Fit3 fits the same model with the response variable that's log transformed, this might be due to the response variable showing possible right skewness in the pairs plot.

c.

AIC: The model chosen by this selection criteria is the model with `public`, `ACT`, `price`, `pct_need` as predictors, the next lowest AIC is the model with those predictors plus `pct_merit_aided`. The difference between these models is 1.9 so these models are essentially equivalent in terms of AIC so we adopt the principle of parsimony and choose the model with the least predictors which is the model with `public`, `ACT`, `price`, `pct_need`.

BIC: The model with the lowest BIC is the model with `public`, `ACT`, `price`, `pct_need`. The next model with the lowest BIC is the one with those 3 predictors with `pct_merit_aided` added however the difference is 6.2 and the rule of thumb states that when the difference is greater than 6 we have a strong preference for the model with the lower BIC which is the model with `public`, `ACT`, `price`, `pct_need`.

d.

y should be log transformed.

e.

```r
exp(0.0093865)-1
```

```
## [1] 0.009430691
```

The estimated effect of `ACT` scores on `earn` is a multiplicative factor of 0.009 when all other predictors are kept constant.

f.

All smooth terms are non-linear and significant.

g.

p-values for the smooth terms for `price` and `pct_need` are high whilst `ACT` is low, however the edfs are all much lower than k' and k-indexes are all high so there's no indication that more basis functions are needed.

h.

The linear model isn't complex enough to explain the variance in the data, linear models assume that relationships between the response and the predictors are linear, GAMs on the other hand can fit linear relationships together with non-linear relationships and is much better at capturing the variance hence by the adjusted r-squared is higher for the GAM.

i.

AIC: Prefers the GAM model, the difference in AIC between the models is 11 so the model with the lower AIC is always preferred.

BIC: Prefers the lm model, the difference is 18 and the rule of thumb states a difference of 10 or more strongly prefers the model with lower BIC.

j.

In this scenario our objective is inference rather than prediction so since the results of the AIC and BIC selection criteria is different, we look towards the adjusted R-squared values and see that the GAM has the higher adjusted R-squared and thus that is the model I would present since in this case model fit is a more important factor than interpretability.

2.

a. FALSE

The lm function only fits a least squares estimation model. glm on the other hand fits a maximum likelihood estimation model.

b. TRUE

-ish. Standardised residuals are much better at being used to detect outliers but raw residuals are much easier to interpret.

c. FALSE

The t-test only test the assertion that it's reasonable that the predictor could be zero. null hypothesis: $H_0 : \beta_2 = 0$ vs. $H_0 : \beta_2 \neq 0$ so all it's saying is that there's no evidence to reject the null hypothesis that the coefficient can be zero.

d. FALSE

In the model, y is the response, s(X1) is X1 added as a smooth term and X2 is fitted as a linear term. Together with the intercept there's only 3 regression coefficients in the model.