HW 2 – ASML

Group-5: Youmin Li, Anastasia, Xiangcheng

ISLR CH3. Question 4:

(a) Training RSS (True Relationship is Linear)

If the true relationship is linear (), the cubic regression model will have a training RSS that is lower than or equal to that of the linear model. Because the cubic model is more flexible (it includes X, X², X³ terms) and the linear model is essentially a special case of the cubic (with β2 = β3 = 0). In least squares, adding predictors cannot increase the minimum training RSS, the extra terms can always be set to zero to mimic the linear fit, or adjusted to capture more of the data’s variation. Therefore, we expect the cubic fit to achieve a smaller training RSS by fitting the data closer.

(b) Test RSS (True Relationship is Linear)

For test data (new observations), we expect the linear regression to have a lower RSS (better prediction error) than the cubic model when the true relationship is linear. The linear model is correctly specified and will generalize well, whereas the cubic model’s extra flexibility will tend to overfit the training noise without providing real signal improvements. Those unnecessary quadratic and cubic terms will fit random quirks in the training set (since the true function has no such curvature), hurting performance on new data. In other words, the cubic model’s lower training RSS comes at the cost of higher variance and worse prediction accuracy. Thus, the linear model should yield a smaller test RSS than the cubic model in this scenario.

(c) Training RSS (True Relationship is Non-linear)

If the true relationship between X and Y is not linear, the cubic regression will still have a training RSS that is lower than or equal to the linear regression. The more the true function deviates from linearity, the more the flexible cubic model can exploit its extra degrees of freedom to fit the training data closely. Regardless of the exact underlying shape, a model with higher flexibility can always fit the observed points at least as well as a less flexible model. So the cubic model is expected to better fit the training data whenever the true pattern isn’t strictly linear, simply because it can capture curvature that a straight line cannot.

(d) Test RSS (True Relationship is Non-linear)

When the true relationship is non-linear, it is not guaranteed which model will have the lower test RSS – we do not have enough information to tell without knowing how “far” from linear the truth is. This outcome depends on the nature and degree of the non-linearity (a bias-variance trade-off situation). If the true function is only mildly non-linear (close to linear), the linear model’s lower complexity (low variance) might outperform the cubic model on test data – the linear model would suffer some bias but also avoids overfitting, potentially yielding smaller test RSS. On the other hand, if the true relationship has pronounced curvature (e.g. closer to a cubic shape or otherwise strongly non-linear), the cubic regression (with its additional flexibility and lower bias for non-linear patterns) is likely to achieve lower test RSS than the linear model. Without knowing how significant the non-linearity is, we cannot definitively choose one model over the other for test error. In summary, it depends on whether the true function is closer to a straight line or requires the curved terms – that will determine whether the simpler linear model or the more flexible cubic model yields the lower test RSS on new data.

ISLR CH3. Question 10:

1. **multiple regression model prediction**

lm(formula = Sales ~ Price + Urban + US, data = Carseats)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 13.043469 0.651012 20.036 < 2e-16 \*\*\*

Price -0.054459 0.005242 -10.389 < 2e-16 \*\*\*

UrbanYes -0.021916 0.271650 -0.081 0.936

USYes 1.200573 0.259042 4.635 4.86e-06 \*\*\*

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.472 on 396 degrees of freedom

Multiple R-squared: 0.2393, Adjusted R-squared: 0.2335

F-statistic: 41.52 on 3 and 396 DF, p-value: < 2.2e-16

1. **Interpretation of coefficients**

Intercept: expected Sales when Price=0, Urban="No", US="No".

Price: the expected change in Sales for a 1 unit increase in Price, holding Urban and US fixed.

UrbanYes: difference in mean Sales between Urban="Yes" and Urban="No", holding Price and US fixed.

USYes: difference in mean Sales between US="Yes" and US="No", holding Price and Urban fixed.

1. **Model formation**

**(d)** We can reject , because p value of these three parameter << 0.05, which means statistically significant with more than 95% confident level.

**(e) reduced model:**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 13.03079 0.63098 20.652 < 2e-16 \*\*\*

Price -0.05448 0.00523 -10.416 < 2e-16 \*\*\*

USYes 1.19964 0.25846 4.641 4.71e-06 \*\*\*

Residual standard error: 2.469 on 397 degrees of freedom;

Multiple R-squared: 0.2393, Adjusted R-squared: 0.2354,

F-statistic: 62.43 on 2 and 397 DF, p-value: < 2.2e-16

**(f) model fitness**

All estimated parameters are statistically significant, so fitness is improved.

**(g) Confidence Interval**

2.5 % 97.5 %

(Intercept) 11.79032020 14.27126531

Price -0.06475984 -0.04419543

USYes 0.69151957 1.70776632

Question 1.1

Question 1.2

Question 1.3

Question 1.4