STAT 523 HW 11

Md Muhtasim Billah

4/21/2020

Question

Madgyal sheep is indigenous breed of sheep and distributed in Maharashtra state of India. Due to high growth rate, this breed is of special importance for growth traits studies. A study records the body weights (lbs) for 60 lambs and their age (month). A simple linear regression model was implemented to analyze the relation between lamb weight and age. The results are shown below.

a. Do the simple linear regression model assumptions appear to be satisfied? Discuss each assumption separately.
b. Regardless of your answer in a. Use the model results to build a 95% prediction interval for a 10-month-old lamb. Use $S_{xx} = 719.345$ and $\bar{x} = 6.333$.
c. Regardless of your answer in a. Use the model results to build a 95% confidence interval for the average weight of lambs that are 10-month-old. Which interval is wider?
d. Can you predict the weight for a 3-year-old lamb? Why or why not?

The analysis was re-run on the log transformed data. The new results are shown below. Use it to answer the following questions e to g.

e. Do the simple linear regression model assumptions appear to be satisfied? Discuss each assumption separately.
f. Discuss the advantage of the log transformation in terms of the linear regression model.
g. Run a hypothesis test to check if there is a significant linear relation between the log of lamb weight and log of lamb age.

Answer

a.

There are three basic assumptions for a simple linear regression model which are linear correlation among the predictor and response variables and normality and equal variance of the residuals.

i) Linear correlation of the variables:

From the scatter plot, it looks like the correlation between the predictor and the response is not very linear, rather their might be a possibility for a curvilinear relationship. Thus, the primary assumption of simple linear regression is possibly violated.

ii) Normality of the residuals:

From the provided normal probability (Q-Q) plot, it looks like though most of the residuals fall onto the line, they deviate a lot at the tails. This indicates to a non-normality of the residuals. Thus, the assumption of normality seems to be violated.

iii) Equal variance of the residuals:

From the residuals vs fitted plot, it looks like there is a curvilinear pattern among the residuals rather than a random scatter which indicates to the violation of the assumption of equal variance.

b.

From the sumary of the current regression model, we get the following parameter estimates.
The intercept, $\beta_0 = 7.9163$.
The slope, $\beta_1 = 1.4396$.
At $x = x^*$, the mean response is,

$$\hat{y} = \beta_0 + \beta_1 x^*$$

Thus, for a 10-month-old lamb ($x^* = 10$), the mean response predicted by the model is,

$$\hat{y} = 7.9163 + 1.4396 \times 10 = 22.3132$$

The formula for prediction interval (PI),

$$\hat{y} \pm t_{\alpha/2,\ n-2} \cdot \sqrt{\hat{\sigma}^2 + s_{\hat{y}}^2}$$

Where,

$\alpha = 0.05$, $n = 60$, $\hat{\sigma}^2 = MSE = 3.981$.
$t_{\alpha/2,\ n-2} = t_{0.05/2,\ 60-2} = t_{0.025,\ 58} = 2.001717$.

And,

$$s_{\hat{y}}^2 = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

Where,
$\bar{x} = 6.633$ and $S_{xx} = 719.345$.

Plugging in the values,

$$s_{\hat{y}}^2 = \sqrt{3.981} \times \sqrt{\frac{1}{60} + \frac{(10 - 6.633)^2}{719.345}} = 0.3593$$

Finally, the 95% prediction interval becomes,

$$22.3132 \pm 2.001717 \times \sqrt{3.981 + 0.3593} = 22.3132 \pm 4.17025 = (18.14295, 26.4833)$$

c.

The formula for confidence interval (CI) is the following,

$$\hat{y} \pm t_{\alpha/2,\ n-2} \cdot s_{\hat{y}}$$

The 95% confidence interval for a 10-month-old lamb,

$$22.3132 \pm 2.001717 \times \sqrt{0.3593} = 22.3132 \pm 1.19986 = (21.11334, 23.51306)$$

From the PI and CI, it is evident that PI is wider due to an extra term in its calculation.

d.

We cannot predict the weight for a 3-year-old lamb becuase it would be an extrapolation since the age (the predictor) would exceed the "scope of the model".

e.

Analysis of the three assumptions of the regression rerun on the log-transformed data.

i) Linear correlation of the variables:

From the scatter plot, it looks like the correlation between the predictor and the response is much more linear than before. Thus, the primary assumption of simple linear regression now holds.

ii) Normality of the residuals:

From the provided normal probability (Q-Q) plot, it seems that most of the residuals fall onto the line, and they don't deviate much at the tails either. This indicates to the normality of the residuals. Thus, the assumption of normality is not violated anymore.

iii) Equal variance of the residuals:

From the residuals vs fitted plot, it looks like there is a random scatter of the residuals which indicates to the equal variance of the residuals. Thus, the third assumption holds true as well.

f.

Advantages of log transformation for SLR:

i) Log transformation often helps make the data more interpretable by making it less skewed.
ii) Often a dataset with a nonlinear pattern can be log transformed to gain a fairly linear trend. SLR is less complicated and thus the regression coefficients can still have simple interpretations.
iii) Log transformation can often corrent the violations of assumptions which is another great advantage.

g.

F-test is a measure for determining the overall significance of a regression model in explaining the relation between the predictors and the response. Based on the F-statistic (from the given summary), a hypothesis test can be performed to check if there is a significant linear relation between the log of lamb weight and log of lamb age.

Null hypothesis, H_0 : There is linear no relationship between the predictor and the response.
Alternative hypothesis, H_a : There is strong linear relationship.
Level of significance, $\alpha = 0.05$.
F-statistic, $F = 405.4$.
p-value = $2.2e^{-16}$.

It is clear that,

$$p - value << \alpha$$

Thus, the null can be rejected. This indicates that there is a highly strong linear relationhip between the response (log of lamb weight) and the predictor (log of lamb age) variable.