Assignment1

|  |  |
| --- | --- |
|  |  |
|  | |

表格

描述已自动生成

图表, 散点图

描述已自动生成

图示

描述已自动生成

表格

低可信度描述已自动生成

图表, 折线图

描述已自动生成

* Residual and predicted plots: The residual plots should be randomly distributed near zero, with no obvious pattern. This shows that the linear hypothesis is basically valid, but there is a slight nonlinear tendency.
* Normality: The Q-Q chart shows that the residuals are roughly distributed along a straight line, indicating that the residuals are close to a normal distribution. The histogram also supports this.
* Homoscedasticity: The scatter plot of the residuals and the predicted values shows that the variance of the residuals is relatively constant over the entire range, but there is slight heteroscedasticity.
* Cook's D: There are no obvious high-leverage points in Cook's D plot, indicating that there are no outliers that seriously affect the model.
* Model goodness of fit: An R-squared value of 0.2723 and an adjusted R-squared value of 0.2523 indicate low variability in model interpretation and may need to consider adding other variables or using more complex models.

Overall, the model assumptions are largely satisfied, but the fit can be improved by adding variables or using nonlinear models.

图示, 示意图

描述已自动生成

图形用户界面, 应用程序

描述已自动生成

表格

描述已自动生成

图表, 直方图

描述已自动生成

Factor A (number of layers) has a significant impact on market value.

Factor B (batch effect) As a random effect, may not show a significant effect (further verification is required to verify that the model estimates are correct).

As can be seen from the overall model output, the analysis provides A multi-factor impact assessment of market value in a mixed model evaluating the fixed effects of factor A and the random effects of factor B.

3

1. **R-squared and Additional Predictors**:
   * R-squared is a measure of the proportion of the variance in the dependent variable that is predictable from the independent variables. When you add more predictor variables to a regression model, the R-squared value can either increase or stay the same. It will never decrease because R-squared is calculated based on the fit of the model, and adding more predictors can't make the fit worse.
2. **Adjusted R-squared**:
   * While adding predictors will not decrease the R-squared value, it can affect the Adjusted R-squared, which accounts for the number of predictors in the model relative to the number of data points. Adjusted R-squared can decrease if the added predictors do not improve the model sufficiently to justify their inclusion, helping to guard against overfitting.
3. **Overfitting**:
   * Including all available predictor variables in the model without considering their relevance or significance can lead to overfitting. Overfitting occurs when the model becomes too complex and starts capturing noise rather than the underlying pattern. This can result in a model that performs well on the training data but poorly on new, unseen data.
4. **Model Parsimony**:
   * A good regression model should be as simple as possible while still capturing the essential relationships in the data. This is known as the principle of parsimony. Including only meaningful predictors helps in creating a model that is easier to interpret and generalizes better to other datasets.
5. **Statistical Significance and Multicollinearity**:
   * Not all predictor variables are statistically significant, and including non-significant predictors can complicate the model without providing real benefits. Additionally, multicollinearity (when predictor variables are highly correlated with each other) can cause issues with the interpretation of the model coefficients.

In summary, while adding more predictors to a regression model cannot decrease R-squared, it is important to carefully consider which predictors to include. Using statistical tests, domain knowledge, and validation techniques helps ensure that the model is both accurate and interpretable.

Appendix

Code1

/\* Step 1: Import the data \*/

data senic;

infile '/home/u64080188/SENIC - HW 1.txt' firstobs=1;

input Id

Length\_of\_stay

Age

Infection\_risk

Rculturing\_r

R\_Xray\_r

Beds\_number

Medical\_school\_affiliation

Region

Average\_daily\_census

Number\_of\_nurses

Avai\_facs\_and\_ser;

run;

/\* Step 2: Create a new dataset with the selected columns \*/

data selected\_data;

set senic;

keep Length\_of\_stay Age Medical\_school\_affiliation Average\_daily\_census;

run;

/\* Step 3: Plot Length\_of\_stay against each of the other selected variables \*/

/\* Plot Length\_of\_stay vs Age \*/

proc sgplot data=selected\_data;

title "Scatter Plot of Length of Stay vs Age";

scatter x=Age y=Length\_of\_stay / markerattrs=(symbol=CircleFilled color=red);

xaxis label="Age";

yaxis label="Length of Stay";

run;

/\* Plot Length\_of\_stay vs Medical\_school\_affiliation \*/

proc sgplot data=selected\_data;

title "Scatter Plot of Length of Stay vs Medical School Affiliation";

scatter x=Medical\_school\_affiliation y=Length\_of\_stay / markerattrs=(symbol=CircleFilled color=blue);

xaxis label="Medical School Affiliation";

yaxis label="Length of Stay";

run;

/\* Plot Length\_of\_stay vs Average\_daily\_census \*/

proc sgplot data=selected\_data;

title "Scatter Plot of Length of Stay vs Average Daily Census";

scatter x=Average\_daily\_census y=Length\_of\_stay / markerattrs=(symbol=CircleFilled color=green);

xaxis label="Average Daily Census";

yaxis label="Length of Stay";

run;

/\* Fit a linear regression model \*/

proc reg data=selected\_data;

model Length\_of\_stay = Age Medical\_school\_affiliation Average\_daily\_census;

output out=reg\_results p=predicted r=residual;

run;

/\* Check model assumptions with diagnostic plots \*/

proc sgplot data=reg\_results;

title "Residuals vs Predicted";

scatter x=predicted y=residual;

refline 0 / axis=y lineattrs=(color=red);

run;

proc univariate data=reg\_results normal;

var residual;

histogram residual / normal;

probplot residual / normal(mu=est sigma=est);

run;

2

data t;

infile '/home/u64080188/CH25PR17.txt';

input score A B C;

run;

proc mixed data=t;

class A B C;

model score = A/residual;

repeated /type=cs subject=C;

lsmeans A/diff=all cl;

run;

proc reg data=t;

model score = A B /influence vif;

run;