Ximan Liu
1935858
DSC 423
HW6

I have completed this work independently. The solutions given are entirely my own work.

1)
F-test:
The null hypothesis indicates that the model with no independent variables fits the data as well as our model.
The alternative hypothesis indicates that our model fits the data better than the intercept-only model.
Usage:
The F-test for linear regression tests whether any of the independent variables in a multiple linear regression model are significant. R-squared tells how well our model fits the data, and the F-test is related to it. If none of our independent variables are statistically significant, the overall F-test is also not statistically significant.

T-test:
The null hypothesis assumes that the true mean difference between the paired samples is zero. Under this model, all observable differences are explained by random variation.
The alternative hypothesis assumes that the true mean difference between the paired samples is not equal to zero.
Usage:
If we determine if there is a significant linear relationship between an independent variable x and a dependent variable y, we use a significance test. We will carry out a t-test for the slope by calculating the p-value and comparing it with the desired significance level. Then, we will find the p-value by first determining the t-value or test statistic. This number is the ratio of the difference between the statistic and the parameter and the standard deviation of the statistic. If our p-value is extreme or more extreme than our significance level, we reject the null hypothesis and have significant evidence to conclude the alternative.

2)
We make four assumptions about the residuals when building a model.
Firstly, linearity. We assume that there exists a linear relationship between the independent variable X and the dependent variable Y. To verify that, we can draw a scatter plot of value X and value Y. If the plot seems like can fall along a straight line like linear pattern and not a curve, then that the assumption is true.
Secondly, independence. We assume that the residuals are independent because we do not want a following continuously pattern on residuals. To verify that, we need to check out residual time series plot, as known as a plot of residuals and time. Most of residual falls within 95% confidence brands around aero, which located +/-2 over the square root of n, n

is the size of sample.
Thirdly, normality. We assume that residuals are normal distributed. To verify that, we use quantile-quantile plot (as known as Q-Q plot). It indicates whether the residuals follow normal distribution. If the points in the plot look like a straight diagonal line, the assumption is approved.
Lastly, homoscedasticity. We assume that residuals have constant variance at every level of X. If residuals do not follow this assumption, then the situation is heteroscedasticity. it is quite bad because it increases the variance of the regression coefficient estimates, but the model does not notice that. To verify assumption, we can draw a scatter plot of fitted value versus residuals. If the residuals become more spread out as the fitted values getting larger, then there is a cone shape showed up as a typical sign of heteroscedasticity.

3)
Once we know the size of residuals, we can start assessing how good our regression fit is. Regression fitness can be measured by R squared and adjusted R squared. R-Squared is the goodness of fit, also known as coefficient of determination. It measures explained variation over total variation, the quality of fit.

4)
Given possible predicts x1 and x2, we should know that the order of the polynomial model is kept as low as possible. Some transformations can be used to keep the model to be of the first order. If this is not satisfactory, then the second-order polynomial is tried. Moreover, the polynomial models can be used in situations where the relationship between explanatory variables is curvilinear. A nonlinear relationship in a small range of explanatory variable can also be modelled by polynomials. It also can be used to approximate a complex nonlinear relationship. However, arbitrary fitting of higher-order polynomials can be a serious abuse of regression analysis. A model which is consistent with the knowledge of data and its environment should be considered.

5)
Beta-0 = -1338.95134
Beta-1 = 12.74057
Beta-2 = 85.95298

Y = -1338.95134 + 12.74057 * AGE + 85.95298 * NUMBIDS + e

The model minimizes the sum of the square of the errors.

Hence the p-value is so low, so we reject the null hypothesis and accept the alternative that beta-1 is not equal to 0 and assume that it equals the prediction of 12.74057. Likewise, the p-

value is so low, so we reject the null hypothesis and accept the alternative that beta-2 is not equal to 0 and assume that it equals the prediction of 85.95298.

Overall, we need to include AGE and NUMBIDS in our regression model.

SSE (Sum of Squares Error) = 516727.

R2 = 0.8923 (89.23%). It indicates our model is predicting 89.23 percent of the variability in our dependent variable Y.

MSE (Mean Squared Error) = 17818. It measures the average of the squares of the errors, that is the average squared difference between the estimated values and the actual value. It corresponds to the expected value of the squared error loss.

RMSE (Root Mean Square Error) = 133.48467. It is the standard deviation of the residuals (prediction errors) and tells how concentrated the data is around the line of best fit. It indicates how big our errors are going to be when we make predictions.

6)
Cross-validation (k-fold cross-validation):
a resampling procedure used to evaluate regression models on a limited data sample. It has a single parameter called k that refers to the number of groups that a given data sample is to be split into.

Leave-one-out cross-validation:
a special case of cross-validation where the number of folds equals the number of instances in the data set. The algorithm is applied once for each instance, using all other instances as a training set, and using the selected instance as a single-item test set.

7)
Adj-R2 is the better model when comparing models that have a different amount of variables. R2 always increases when the number of variables increases. Even if adding a useless variable to model, R2will still increase. We should always compare models with different number of independent variables with Adj-R2. Adj-R2 only increases if the new variable improves the model more than would be expected by chance.

8)
Parsimonious models are simple models with great explanatory predictive power. They explain data with a minimum number of parameters, or predictor variables, and have optimal parsimony, or just the right amount of predictors needed to explain the model well.

A model has few parameters but achieves a satisfactory level of goodness of fit should be preferred over a model that has a ton of parameters and achieves only a slightly higher level of goodness of fit. It is because models with fewer parameters are easier to interpret and understand; also tend to have more predictive ability and perform better when applied to new data.

9)

Categorical variables (qualitative variables) are the variables classify observations into groups. They have a limited number of different values as known as levels. In that case, they cannot be entered into the regression equation directly. Instead, categorical variables need to be recoded into a series of variables. They are recoded into a set of separate binary variables. The procedure is called "dummy" and leads to the creation of a table called contrast matrix.

For example, we assume the regression model is y = b0 + b1*x. The gender of individuals is a categorical variable that can take two levels: male (-1) or female (+1). Then, (b0 - b1) if person is male, (b0 + b1) if person is female. In that case, gender level situations determine the regression coefficient.

10)

Forward stepwise regression:

Pros:

1) suitable scenarios - if the number of variables under consideration is larger than the sample size.
2) do not have to consider the full model.
3) easy to run in statistical packages.
4) runs fast.

Cons:

1) considering collinearity (when variables in a model are correlated which each other), none of them might be kept in the model.
2) does not consider all possible combination of potential predictors.
3) the regression coefficients, confidence intervals, p-values and R2 outputted may be biased and cannot be trusted.
4) the selection of variables will be highly unstable when having a small sample size compared to the number of variables.

Backward stepwise regression:

Pros:

1) start with a full model can consider the effects of all variables simultaneously.
2) considering collinearity (when variables in a model are correlated which each other), backward stepwise may keep them all in the model.

3) easy to run in statistical packages.
4) runs fast.

Cons:
1) not a good choice if the number of candidate variables > sample size (or number of events).
2) does not consider all possible combination of potential predictors.
3) the regression coefficients, confidence intervals, p-values and R2 outputted may be biased and cannot be trusted.
4) the selection of variables will be highly unstable when having a small sample size compared to the number of variables.

All-possible regression:
Pros:
1) fits all possible models based on the independent variables.
2) the number of models fits multiplies quickly.
3) displays the best fitting models of different sizes up to the full model.
Cons:
1) runs slow and time cost may be too high.