Stat 6021 R Tutorial: Multiple Linear Regression

For this tutorial, we will see how to fit a multiple linear regression model in R. We will use the "delivery.txt" data set, from Example 3.1 in the textbook. A soft drink bottler is analyzing the vending machine service routes in his distribution system. He is interested in predicting the amount of time required by the route driver to service the vending machines in an outlet. The industrial engineer responsible for the study has suggested that the two most important variables affecting the delivery time (y) are the number of cases of product stocked (x_1) and the distance walked by the route driver (x_2). The engineer has collected 25 observations.

1. Fit a multiple regression model using *Delivery* as the response and the other variables as predictors. To use `lm()` for multiple regression, type

   ```
   result<-lm(Delivery~Number+Distance)
   ```

   Each predictor in `lm()` is appended after a + sign.

2. Critical values. Another approach to decide whether to reject the null hypothesis is using the critical value (instead of the p-value). The critical value is the smallest (in magnitude) value of the test statistic that leads to rejecting the null hypothesis. Another way to view the critical value is that the p-value associated with the critical value is the value of the significance level, α. To find the critical value of an F distribution, use `qf(p, df1, df2)` where p is $1 - \alpha$, and df1 and df2 are the numerator and denominator degrees of freedom respectively. For this example, type `qf(0.95, 2, 22)`. You will notice that our test statistic for the ANOVA F test is larger than this critical value, hence we reject the null hypothesis. Not surprisingly, the p-value is less than 0.05.

3. Confidence interval for regression coefficients. The `confint()` function obtains the confidence intervals for the regression parameters. For example

   ```
   confint(result,level = 0.95)
   ```

4. Confidence interval for mean response, prediction interval for the response of a future observation. The `predict.lm()` function helps with obtaining fitted values for given values of the predictors, and the corresponding confidence and prediction intervals. Suppose for *Number*= 20 and *Distance*= 200, we want the corresponding CI for mean *Delivery*. The following code can be used:

```
newdata<-data.frame(Number=20, Distance=200)
predict.lm(result, newdata, level=0.95, interval="confidence")
predict.lm(result, newdata, level=0.95, interval="prediction")
```

5. By now, you may realize using R for multiple linear regression is very similar to using R for simple linear regression (you can compare this tutorial with the tutorials from modules 1 and 2). In fact, the way to generate the residual plot, ACF plot, and QQ plot is also similar. Can you generate these for this data set (with or without looking at the tutorial from module 3)?