Stat509 Fall 2014 Homework 8

Instructor: Peijie Hou

11/18/2014

Instruction: Quiz based on this homework will be given on 12/02/2014. Have a good Thanksgiving!

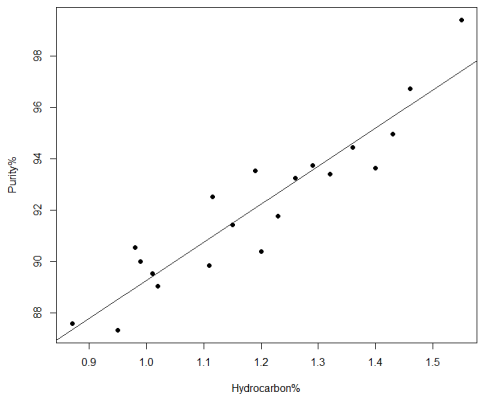
1. We will look at the purity of oxygen (in percent) produced in a chemical distillation process and the percentage hydrocarbons that are present in the main condenser of the distillation unit. The purity of the oxygen will be the response while the percentage of hydrocarbons will be the regressor. The data is given by

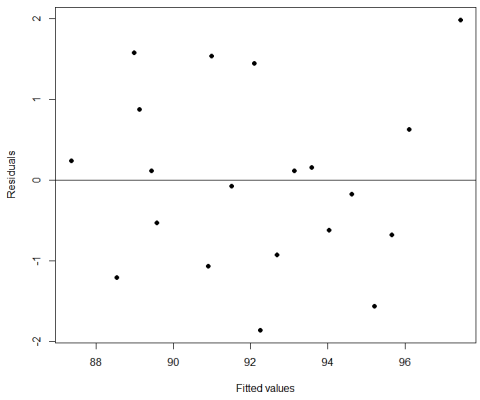
```
Hydrocarbon%:  7.78 0.99 1.02 1.15 1.29 1.46 1.36 0.87 1.23 1.55 1.4 1.19 1.115 0.98 1.01 1.11 1.2 1.26 1.32
1.43 0.95
Purity%:  4.54 90.01 89.05 91.43 93.74 96.73 94.45 87.59 91.77 99.42 93.65 93.54 92.52 90.56 89.54 89.85 90.39
93.25 93.41 94.98 87.33
```

   You can use the following R code:

```
hydrocarbon<-c(0.99,1.02,1.15,1.29,1.46,1.36,0.87,1.23,1.55,1.4,1.19,1.115,0.98,1.01,1.11,1.2,1.26,1.32,
1.43,0.95)
purity<-c(90.01,89.05,91.43,93.74,96.73,94.45,87.59,91.77,99.42,93.65,93.54,92.52,90.56,89.54,89.85,90.39,
93.25,93.41,94.98,87.33)
fit = lm(purity~hydrocarbon)
#scatterplot with regression line superimposed
plot(hydrocarbon,purity,xlab = "Hydrocarbon%",ylab = "Purity%",pch=16)
abline(fit)
#residual plot
# Residual plot
plot(fitted(fit),residuals(fit),pch=16,
xlab="Fitted values",ylab="Residuals")
abline(h=0)
#QQ plot
resid<-residuals(fit)
qqnorm(resid);qqline(resid)
#Find coefficient estimate
summary(fit)
#ANOVA table
anova(fit)
```

   (a) Run the regression in R. Superimpose a fitted regression line on the scatter plot of Hydrocarbon% versus Purity%. Plot residuals vs x values, which assumption does this plot check? Plot QQ plot for residuals, which assumption does this plot check? Are the assumptions met?
   Solution:





There is no pattern in the residual plot (not shown), which indicates the error term has mean 0 and constant variance. The QQ plot check the normality of error terms. The QQ plot (not shown) indicates that it is reasonable to assume the error terms follows a normal distribution.

(b) Identify the sum of squares for the model from the ANOVA table. What does it measure?
Solution:

```
Analysis of Variance Table

Response: purity
            Df  Sum Sq Mean Sq F value    Pr(>F)
hydrocarbon  1 150.780 150.780  120.11 2.142e-09 ***
Residuals   18  22.597   1.255
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The sum of squres for the model is 150.78, it measures the total variability described by the model.

(c) Conduct a hypothesis test for testing $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$. Calculate a 95% confidence interval for β_1, interpret your result.
Solution: The p-value of the test is almost 0 (2.14e-09), so we reject H_0 and conclude that $\beta_1 \neq 0$. The confidence interval of β is

$$\hat{\beta} \pm t_{n-2}\widehat{s.e.(\hat{\beta})} = 14.833 \pm 2.10(1.353) = (11.99, 17.67).$$

We are 95% confident that the true value of β is covered by above interval.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   74.446      1.636   45.51  < 2e-16 ***
hydrocarbon   14.833      1.353   10.96 2.14e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(d) Write an interpretation of the estimated coefficient $\hat{\beta}_1$.
Solution: The mean purity is estimated to increase 14.833 percent with 1 percent increase of hydrocarbones.

(e) Write an interpretation of the coefficient $\hat{\beta}_0$.
Solution: The mean purity is estimated to be 74.446% when percent of hydrocarbon is 0.

(f) What is the coefficient of correlation between hydrocarbon%(x) and purity% (Y)? What is the percentage of total sample variation explained by linear regression?
Solution:

```
> cor(hydrocarbon,purity)
[1] 0.93256
```

The coefficient of correlation is 0.93. Alternatively, you can check the correlation of determination and take the square root of it. Here, $r^2 = 0.8697$ (Multiple R-squared), so $r = \sqrt{0.8697} = 0.93$. We choose positive square root since x and y are positively correlated by checking the scatter plot. The percentage of total sample variation explained by linear regression equals to correlation of determination. The regression model explains about 87% of the total variation.

```
Residual standard error: 1.12 on 18 degrees of freedom
Multiple R-squared:  0.8697,    Adjusted R-squared:  0.8624
F-statistic: 120.1 on 1 and 18 DF,  p-value: 2.142e-09
```

(g) Predict the mean percentage of oxygen purity when 1.3% hydrocarbons is used. Calculate a 95% confidence interval for that predicted mean response by R, interpret your result.
Solution:

```
>predict(fit,data.frame(hydrocarbon=1.3),level=0.95,interval="confidence")
       fit      lwr      upr
1 93.72914 93.12294 94.33534
```

The predicted mean response is 93.73 when hydrocarbon percentage is 1.3 , and we are 95% confident that the mean purity when hydrocarbon percentage is 1.3 is between 93.12 and 94.33.

(h) Calculate a 95% prediction interval for a future observation when hydrocarbons percent is 1.3% by R, interpret your result.
Solution:

```
> predict(fit,data.frame(hydrocarbon=1.3),level=0.95,interval="prediction")
       fit     lwr      upr
1 93.72914 91.2984 96.15988
```

We are 95% confident that a future observation is between 91.3 and 96.16 when hydrocarbons percent is 1.3%.