Lab 9

David Wiley / Duy Truong

February 27, 2019

Problem 3.11:

Parts a-d. Write down the estimated model. Interpret in the context of the problem whenever necessary.

An engineer performed an experiment to determine the effect of CO_2 pressure, CO_2 temperature, peanut moisture, CO_2 flow rate, and peanut particle size on the total yield of oil per batch of peanuts. Table B.7 summarizes the experimental results.

a.

Fit a multiple linear regression model relating yield to these regressors.

```
# Reading in the data
dat = read.csv("C:\\Users\\Nick\\Documents\\0_Spring 2019\\Applied Regression\\Labs_HW\\Data_Sets\\B7.cs

y = dat$y
x1 = dat$x1
x2 = dat$x2
x3 = dat$x3
x4 = dat$x4
x5 = dat$x5
n = length(y)
p = length(dat[1,]) - 1

# Fitting the data into a table
fit = lm(y~., dat)
sumfit = summary(fit)

coeff = round(coef(fit), digits = 3)
B0H = getElement(coeff, "(Intercept)")
B1H = getElement(coeff, "x1")
B2H = getElement(coeff, "x2")
B3H = getElement(coeff, "x3")
B4H = getElement(coeff, "x4")
B5H = getElement(coeff, "x5")


sumfit
```

```
##
## Call:
## lm(formula = y ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.250  -4.438   0.125   5.250   9.500
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.208e+01  1.889e+01   2.757 0.020218 *
## x1           5.556e-02  2.987e-02   1.860 0.092544 .
## x2           2.821e-01  5.761e-02   4.897 0.000625 ***
## x3           1.250e-01  4.033e-01   0.310 0.762949
## x4          -8.990e-17  2.016e-01   0.000 1.000000
## x5          -1.606e+01  1.456e+00 -11.035  6.4e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.065 on 10 degrees of freedom
## Multiple R-squared:  0.9372, Adjusted R-squared:  0.9058
## F-statistic: 29.86 on 5 and 10 DF,  p-value: 1.055e-05
```

The MLR model is:

$$\hat{y} = (52.079) + (0.056)x_1 + (0.282)x_2 + (0.125)x_3 + (0)x_4 + (-16.065)x_5$$

b.

Test for significance of regression. What conclusions can you draw?

$H_0 : \hat{\beta_0} = H_0 : \hat{\beta_1} = ... = H_0 : \hat{\beta_j} = 0$;

$H_1 : \hat{\beta_j} \neq 0$; for any j

```
fitted = anova(fit)
fitted
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## x1         1  225.0   225.0   3.4589 0.0925445 .
## x2         1 1560.2  1560.2  23.9854 0.0006254 ***
## x3         1    6.2     6.2   0.0961 0.7629488
## x4         1    0.0     0.0   0.0000 1.0000000
## x5         1 7921.0  7921.0 121.7679 6.401e-07 ***
## Residuals 10  650.5    65.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SSr = sum(fitted[1:5,2])
SSres = fitted[6,2]
SSt = sum(getElement(fitted, "Sum Sq"))


F0 = sumfit$fstatistic[1]
# Long way: F0 = (SSr/p) / (SSres / (n-p-1))
F0
```

```
##    value
## 29.86164
```

```
pvalue = 1 - pf(F0, p, n-p-1)
pvalue
```

```
##        value
## 1.054723e-05
```

Since the p-value is less than α, we reject H_0 and conclude that there is, in fact, a linear relationship between the regressors.

c.

Use t-tests to assess the contribution of each regressor to the model. Discuss your findings.

```
t_coefs = round(sumfit$coefficients[2:6,"t value"], digits = 3)

# Critical t-value
t_crit = round(qt(.975,n-p-1), digits = 3)
```

Our critical t-value to compare to is 2.228. Our t-values for each regressor is: $1.86, 4.897, 0.31, 0, -11.035$

Therefore, we can see we will reject $H_0 : \hat{\beta}_1 = 0$, $H_0 : \hat{\beta}_2 = 0$, $H_0 : \hat{\beta}_3 = 0$, $H_0 : \hat{\beta}_5 = 0$. But we do not reject $H_0 : \hat{\beta}_4 = 0$ since it does equal 0. This means $H_0 : \hat{\beta}_4 = 0$ has 0 contribution to our model.

d.

Calculate R^2 and R^2_{ADJ} for this model. Compare these values to the R^2 and R^2_{ADJ} for the multiple linear regression model relating yield to temperature and particle size. Discuss your results.

```
# R^2 and R^2 adjusted for the model
r_sq = round(sumfit$r.squared, digits = 3)
r_sq_adj = round(sumfit$adj.r.squared, digits = 3)


# R^2 and R^2 adjusted for CO2 temperature (x2) and particle size (x5)
sumfit_tps = summary(lm(y~x2+x5, dat))


r_sq_tps = round(sumfit_tps$r.squared, digits = 3)
r_sq_tps_adj = round(sumfit_tps$adj.r.squared, digits = 3)

sumfit_tps
```

```
##
## Call:
## lm(formula = y ~ x2 + x5, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.375  -4.188  -0.875   3.438  12.625
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  80.13461    5.69146  14.080 3.01e-09 ***
## x2            0.28214    0.05883   4.796 0.000349 ***
## x5          -16.06498    1.48659 -10.807 7.26e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.236 on 13 degrees of freedom
## Multiple R-squared:  0.9149, Adjusted R-squared:  0.9018
```

```
## F-statistic: 69.89 on 2 and 13 DF,  p-value: 1.107e-07
```

From the summary table we can see that $R^2 = 0.937$ for the entire model, which means that 93.7% of the variation in y can be explained by the model. Where as $R^2 = 0.915$ for the temperature and particle size regressors. For the entire model, $R^2_{ADJ} = 0.906$, which means that 90.6% of the variation can be explained by the independent variables in the model. Where as, $R^2_{ADJ} = `rr_sq_tps_adj$ for the temperature and particle size regressors.

Both, the R^2 and R^2_{ADJ} for just temperature and particle size, are smaller than for the entire model. This is reasonable because we are missing the other regressors in the model to account for the variation. The more regressors we include in the model from the data, NOT adding more regressors, the more variation we would be able to account for.

e.

Find a 95% CI for the regression coefficient for temperature for both models in part d. Discuss any differences.

```
#confint(fit, level = .95)

#confint(lm(y~x2+x5, dat), level = .95)

temp_CI_1 = c(B2H-t_crit*sumfit$coefficients[3, "Std. Error"], B2H+ t_crit*sumfit$coefficients[3, "Std.

temp_CI_2 = c(sumfit_tps$coefficients[2]-t_crit*sumfit_tps$coefficients[2, "Std. Error"], sumfit_tps$co
```

The CI for temperature in the first model is:

$$0.1536456, 0.4103544$$

The CI for temperature in the second model is:

$$0.1510774, 0.4132083$$

The CI for temperature in the model limited to two regressors is larger than those in the model that includes all of the regressors from the data. This is because having less regressors that affect the model provides a less accurate CI which means we have to account for a little more variability in the values we accept. Since we have more regressors in the original model that we take into consideration, we have a smaller, more accurate confidence interval.