Regression Analysis - STAT 482 - Exam 1: Due December 2, 2021

Alex Towell (atowell@siue.edu)

Problem 1

Experience with a certain type of plastic indicates that a relationship exists between the hardness (measured in Brinell units) of items molded from the plastic (response variable y) and the elapsed time (measured in hours) since the termination of the molding process (input variable x). Sixteen batches of the plastic were made, and from each batch one test item was molded. Each test item was randomly assigned to one of the four predetermined time levels. The data is available on Blackboard as a csv file.

Part (a)

State the simple linear regression model.

Reproduce

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where

$$\begin{aligned} \epsilon_1, \ldots, \epsilon_n &\sim \mathcal{N}(0, \sigma^2) \\ Y_i &= \text{is the } i\text{-th plastic hardness response (in Brinell units)} \\ x_i &= \text{is the } i\text{-th elapsed molding time.} \end{aligned}$$

Part (b)

Provide an interpretation for the regression parameter β_1, stated in the context of the problem.

Reproduce

β_1 is the difference in mean plastic hardness (in Brinell units) from a 1 hour increase in elapsed molding time.

Part (c)

Compute an interval estimate for β_1.

Output

```
dat = read.csv("./exam1-1.csv")
dat
```

```
##    time hardness
## 1    16      195
## 2    16      203
## 3    16      196
## 4    16      206
## 5    24      223
## 6    24      217
## 7    24      217
## 8    24      209
## 9    32      232
## 10   32      235
## 11   32      224
## 12   32      233
```

```
## 13   40      257
## 14   40      249
## 15   40      250
## 16   40      249
```

```
#dat$time = as.double(dat$time)
reg.mod = lm(hardness ~ time, data=dat)
confint(reg.mod)
```

```
##                  2.5 %     97.5 %
## (Intercept) 157.299114 174.300886
## time          1.813919   2.392331
```

Reproduce

A 95% confidence interval for β_1 is $[1.814, 2.392]$.

Part (d)

Compute a confidence interval for μ_h at $x_h = 40$ hours.

Output

```
x.h = 40
predict(reg.mod,data.frame(time=x.h),interval="confidence")
```

```
##       fit      lwr      upr
## 1 249.925 245.5966 254.2534
```

Part (e)

Compute a prediction interval for $Y_{h(\text{new})}$ at $x_h = 40$ hours.

Output

```
predict(reg.mod,data.frame(time=x.h),interval="predict")
```

```
##       fit      lwr      upr
## 1 249.925 238.7092 261.1408
```

Part (f)

Explain the difference between a confidence interval and a prediction interval, stated in the context of the problem.

Reproduce

A confidence interval is for the mean plastic hardness (in Brinell units) for all items molded with an elapsed time of 40 hours.

A prediction interval is for the mean plastic hardness (in Brinell units) for a particular item molded with an elapsed time of 40 hours.

Part (g)

State the equations for $E(\text{MS}_\text{R})$ and MS_E.

Reproduce

$$E(\text{MS}_\text{R}) = \sigma^2 + \text{SS}_\text{X}\beta_1^2$$
$$E(\text{MS}_\text{E}) = \sigma^2.$$

Part (h)

Use part (g) to explain a motivation behind the F test for input effects.

Reproduce

Note that

$$F^* = \frac{\text{MS}_\text{R}}{\text{MS}_\text{E}}.$$

so if $\beta_1 \approx 0$, then $E(\text{MS}_\text{R}) = E(\text{MS}_\text{E})$. Small F^* indicates data compatible with the null model. Large F^* indicates data supporting the alternative model.

Part (i)

Compute the F^* statistic and the p-value. Provide an interpretation of the result, stated in the context of the problem.

Output

```
anova(reg.mod)
```

```
## Analysis of Variance Table
##
## Response: hardness
##           Df Sum Sq Mean Sq F value    Pr(>F)
## time       1 5661.6  5661.6  243.27 3.033e-10 ***
## Residuals 14  325.8    23.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Reproduce

Since p-value = .000, the data is not compatible with the no effect model. We accept the model which includes elapsed molding time as a predictor of hardness.

Part (j)

Compute the coefficient of determination r^2. Provide an interpretation of the result, stated in the context of the problem.

Output

```
summary(reg.mod)$r.squared
```

```
## [1] 0.9455819
```

Reproduce

We estimate that 95% of the variation in hardness is explained by elapsed molding.

Part (k)

State the full model (saturated model) in testing for regression model fit.

Reproduce

There are $c = 4$ levels of the input (predictor) variable x (time) where the i-th level of x has $n_i = 4$ replications for $i = 1, \ldots, c$.

The saturated model is given by

$$Y_{ij} = \mu_i + \epsilon_{ij} \begin{cases} i = 1, \ldots, c \\ j = 1, \ldots, n_i. \end{cases}$$

where

$$\begin{aligned} \epsilon_{ij} &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \\ \mu_i &= \text{mean of response at the } i\text{-th level of } x \\ Y_{ij} &= j\text{-th response at the } i\text{-th level of } x. \end{aligned}$$

Part (l)

Compute SS_{PE} and SS_{LF} for testing the fit of the linear regression model.

Output

```
dat$fact.time = as.factor(dat$time)

sat.mod = lm(hardness ~ fact.time - 1, data=dat)
anova(reg.mod,sat.mod)
```

```
## Analysis of Variance Table
##
## Model 1: hardness ~ time
## Model 2: hardness ~ fact.time - 1
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     14 325.82
## 2     12 299.75  2    26.075 0.5219 0.6062
```

Reproduce

From the output, we see that $\text{SS}_{\text{PE}} = 299.75$ and $\text{SS}_{\text{LF}} = 26.075$.

Part (m)

Compute the F_{LF} statistic and the p-value. Provide an interpretation of the result, stated in the context of the problem.

Output

See output for part (l).

Reproduce

From the output, we see that

$$F_{LF} = \frac{\text{SS}_{\text{LF}}/(c-2)}{\text{SS}_{\text{PE}}/(n-c)} = \frac{\text{SS}_{\text{LF}}/(4-2)}{\text{SS}_{\text{PE}}/(16-4)} = \frac{26.075/2}{299.75/12} = .5219$$

which has a p-value $= .606$.

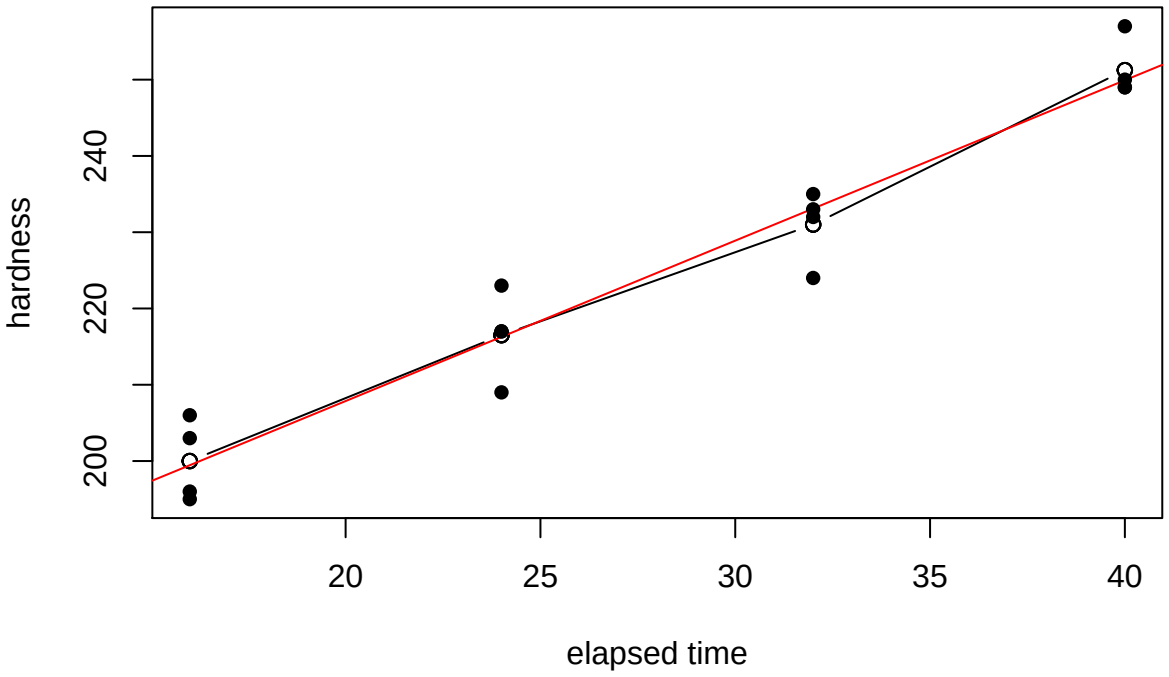
Since the p-value is so large (small F_{LF} value), the experiment finds that the data is compatible with the null model (linear regression model).

Part (n)

Create a plot comparing the fitted values from the regression model with the fitted values from the saturated model.

Output

```
plot(sort(dat$time),
     sat.mod$fitted.values[order(dat$time)],
     type = "b",
     ylab = "hardness",
     xlab = "elapsed time",
     ylim = c(min(dat$hardness),max(dat$hardness)))
abline(reg.mod$coefficients,col="red")
points(dat$time,dat$hardness,pch=16)
```



Part (o)

Explain an advantage of using the regression estimate $\hat{\mu}_{40}$ instead of the sample mean $\bar{y}_{40}$ and explain when this advantage will lead to a more accurate estimator.

Reproduce

Advantage of $\hat{\mu}_{40}$ over $\bar{y}_{40}$: The linear regression estimate of a mean will have smaller variance than the saturated model estimate. When more accurate: If the linear regression model is appropriate, then its estimate $\hat{\mu}_{40}$ will also have a small bias.

Problem 2

Measurements were made on men involved in a physical fitness course. The input variables under consideration are age (in years), weight (in kgs), time to run 1.5 miles (in minutes), heart rate while resting (in bpm), heart rate while running, and maximum heart rate. The goal is to determine which input variables are needed to model the oxygen uptake rate (ml/kg body weight per minute). The data is available on Blackboard as a csv file.

Part (a)

Describe the goal of the discrepancy function approach to model selection. How is the best model defined? What are the two sources of model error?

The goal of model selection is to choose the "best" model from a candidate class of models.

We consider the best model that which provides the most accurate estimation of $E(Y_i|x_i) = \mu_i$ at the observed input levels $\underline{x}_1, \cdots, \underline{x}_n$.

The two sources of model error are model misspecification (bias) and parameter estimation (variance).

Part (b)

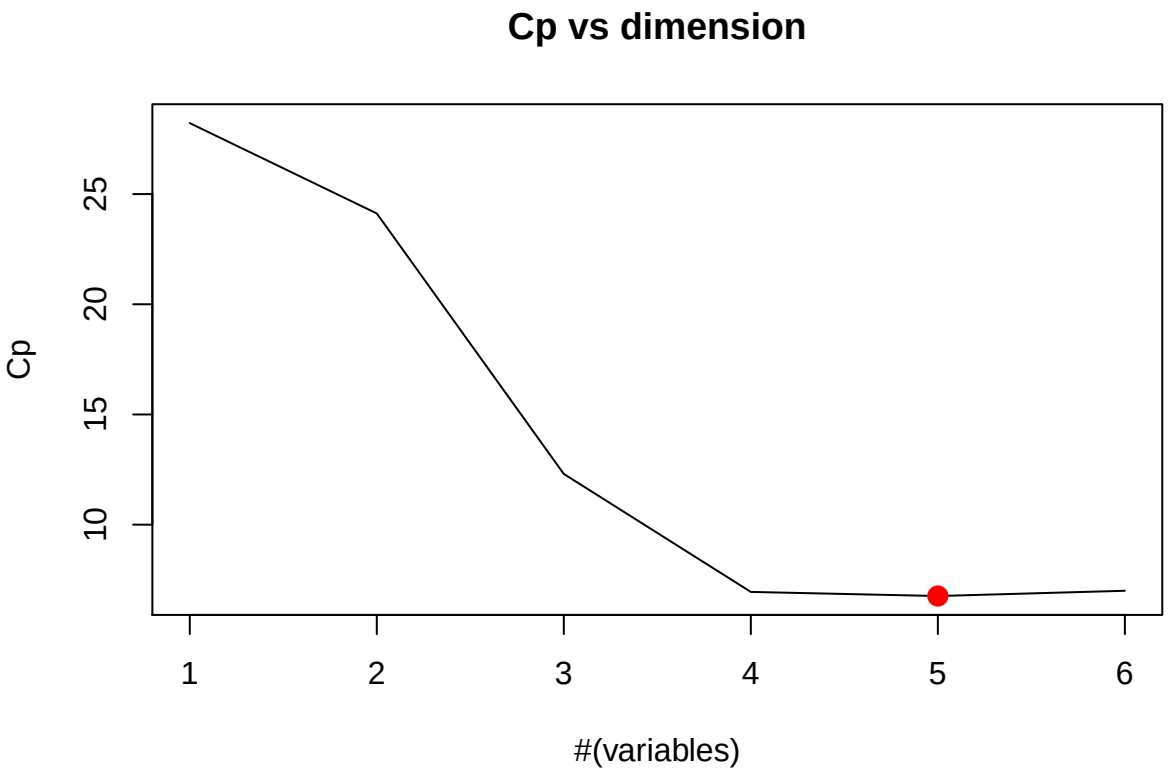
Plot the C_p statistic against the model dimension. Plot the C_p statistic against the candidate models. Which variables are included in the selected model?

Output

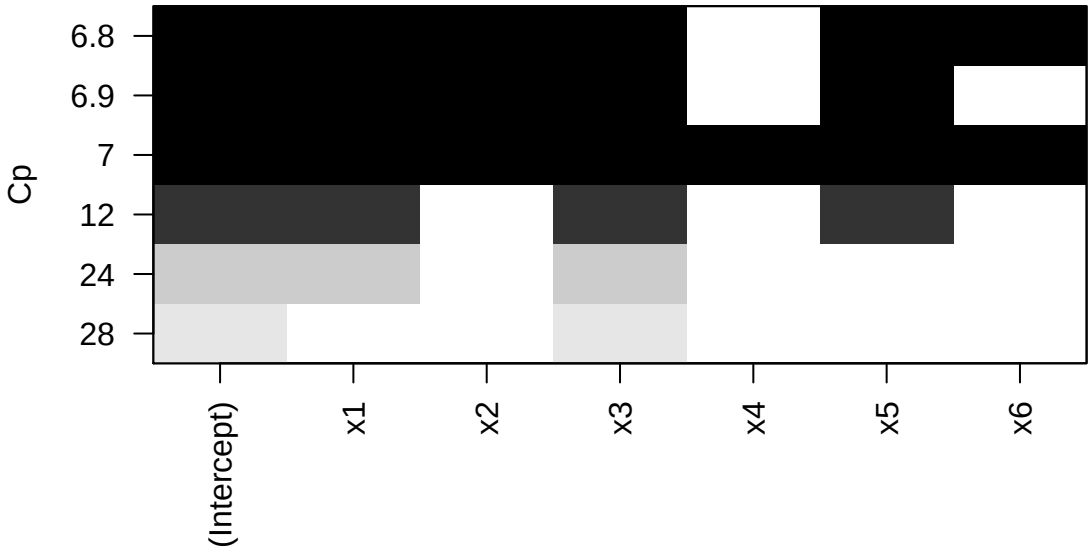
```
library(leaps)
dat2 = read.csv("./exam1-2.csv")
colnames(dat2) = c("x1","x2","x3","x4","x5","x6","y")
head(dat2)
```

```
##   x1    x2    x3 x4  x5  x6        y
## 1 44 89.47 11.37 62 178 182 41.05617
## 2 40 75.07 10.07 62 185 185 48.68738
## 3 44 85.84  8.65 45 156 168 57.72057
## 4 42 68.15  8.17 40 166 172 59.51438
## 5 38 89.02  9.22 55 178 180 52.34193
## 6 47 77.45 11.63 58 176 176 45.30553
```

```
full.mod = regsubsets(y ~ ., data=dat2)
sel = summary(full.mod)
plot(sel$cp,xlab="#(variables)",ylab="Cp",type="l",main= "Cp vs dimension")
star = which.min(sel$cp)
points(star,sel$cp[star],col="red",pch=20,cex=2)
```




```
plot(full.mod,scale="Cp")
```



Reproduce

We select the model which includes 5 input variables,

$$
\begin{aligned}
x_1 &= \text{age},\\
x_2 &= \text{weight},\\
x_3 &= \text{run time},\\
x_5 &= \text{run pulse},\\
x_6 &= \text{max pulse}.
\end{aligned}
$$

Part (c)

Test the selected model against its best competitor having fewer parameters. (Compute the test statistic and the p-value.) How does discrepancy based model selection compare to p-value based selection?

Output

```
m.R = lm(y ~ x1+x2+x3+x5, data=dat2)
m.F = lm(y ~ x1+x2+x3+x5+x6, data=dat2)
anova(m.R,m.F)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2 + x3 + x5
## Model 2: y ~ x1 + x2 + x3 + x5 + x6
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 148.08
## 2     25 136.50  1    11.582 2.1211 0.1577
```

Reproduce

The next best model discards x_6 (max pulse). So, we are testing

$$H_0 : \beta_6 = 0$$

by comparing the model m_F that includes x_6 against the reduced model m_R that discards x_6. See the ANOVA output.

We see that $F^* = 2.121$ with a p-value $= .158$.

We see that in the hypothesis test, we find the data to be compatible with the reduced model, since the F^* statistic has a large p-value. So, according to this test, we would choose m_R. In the discrepancy-based model selection, we found the best model to be m_F. Thus, the rule for adding predictors to a model is less stringent for discrepancy based model selection than for hypothesis testing.