Regression Analysis - STAT 482 - Probem Set 9

Alex Towell (atowell@siue.edu)

Problem 1

> A router is used to cut locating notches on a circuit board. The vibration level is considered to be an important characteristic of the process. Two factors are thought to affect vibration (y): bit size (x_1) and cutting speed (x_2). The data is available as a csv file posted on Blackboard.

Part (a)

> Provide a definition for an interaction effect.

The interaction effect measures the change in an input effect as the other input changes levels.

Part (b)

> Fit an interaction model using coded variables. Compute the coefficient estimates and the standard error.

Coded data transformation

```r
data = read.csv('hw9-1.csv')

to_binary_coded = function(x)
{
  2*(x-mean(x)) / (max(x)-min(x))
}
data$x1 = to_binary_coded(data$bit.size)
data$x2 = to_binary_coded(data$cutting.speed)
data$y = data$vibration
head(data)
```

bit.size	cutting.speed	vibration	x1	x2	y
0.0625	40	18.2	-1	-1	18.2
0.1250	40	27.2	1	-1	27.2
0.0625	90	15.9	-1	1	15.9
0.1250	90	41.0	1	1	41.0
0.0625	40	18.9	-1	-1	18.9
0.1250	40	24.0	1	-1	24.0

Estimates

```r
x.mod = lm(y ~ x1*x2,data=data) # equiv to y~x1+x2+x1:x2 and y~x1+x2+I(x1*x2)
summary(x.mod)
```

```
##
## Call:
## lm(formula = y ~ x1 * x2, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.975 -1.550 -0.200  1.256  3.625
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23.8313     0.6112  38.991 5.22e-14 ***
## x1            8.3188     0.6112  13.611 1.17e-08 ***
## x2            3.7687     0.6112   6.166 4.83e-05 ***
## x1:x2         4.3563     0.6112   7.127 1.20e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.445 on 12 degrees of freedom
## Multiple R-squared:  0.9581, Adjusted R-squared:  0.9476
## F-statistic: 91.36 on 3 and 12 DF,  p-value: 1.569e-08
```

```
coef(x.mod)
```

```
## (Intercept)          x1          x2       x1:x2
##    23.83125     8.31875     3.76875     4.35625
```

We see that

$$\hat{E}(Y|x) = 23.83 + 8.32x_1 + 3.77x_2 + 4.36x_1x_2.$$

Part (c)

Write the estimated regression as a function of x_1 for $x_2 = 1, 0, -1$.

We have the regression function $\hat{E}(Y|x)$, so, for instance, $\hat{E}(Y|x_2 = 0) = 23.83 + 8.32x_1 + 3.77(0) + 4.36(0)x_1 = 23.83 + 8.32x_1$. However, in R, we may compute the slope and intercept estimates with:

```
b0 = coef(x.mod)[1]
b1 = coef(x.mod)[2]
b2 = coef(x.mod)[3]
b12 = coef(x.mod)[4]
reg.estimates.x1 = matrix(c(b0+b2,b0,b0-b2,b1+b12,b1,b1-b12),nrow=3)
dimnames(reg.estimates.x1) = list(c("x2=+1","x2=0","x2=-1"),
                                  c("intercept","slope"))
reg.estimates.x1
```

```
##       intercept    slope
## x2=+1  27.60000 12.67500
## x2=0   23.83125  8.31875
## x2=-1  20.06250  3.96250
```

Thus,

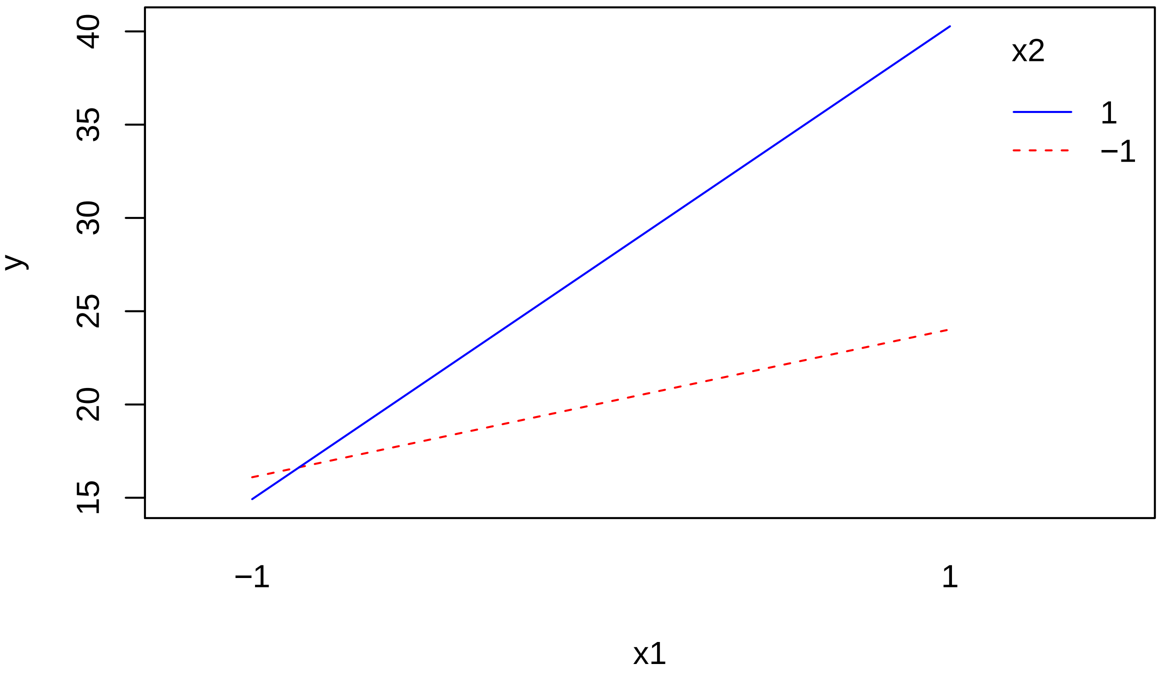
$$\hat{E}(Y|x_1, x_2 = +1) = 27.600 + 12.675x_1,$$
$$\hat{E}(Y|x_1, x_2 = 0) = 23.831 + 8.319x_1,$$
$$\hat{E}(Y|x_1, x_2 = -1) = 20.063 + 3.963x_1.$$

Part (d)

Create interaction plots for both the interaction model and the additive effects model.

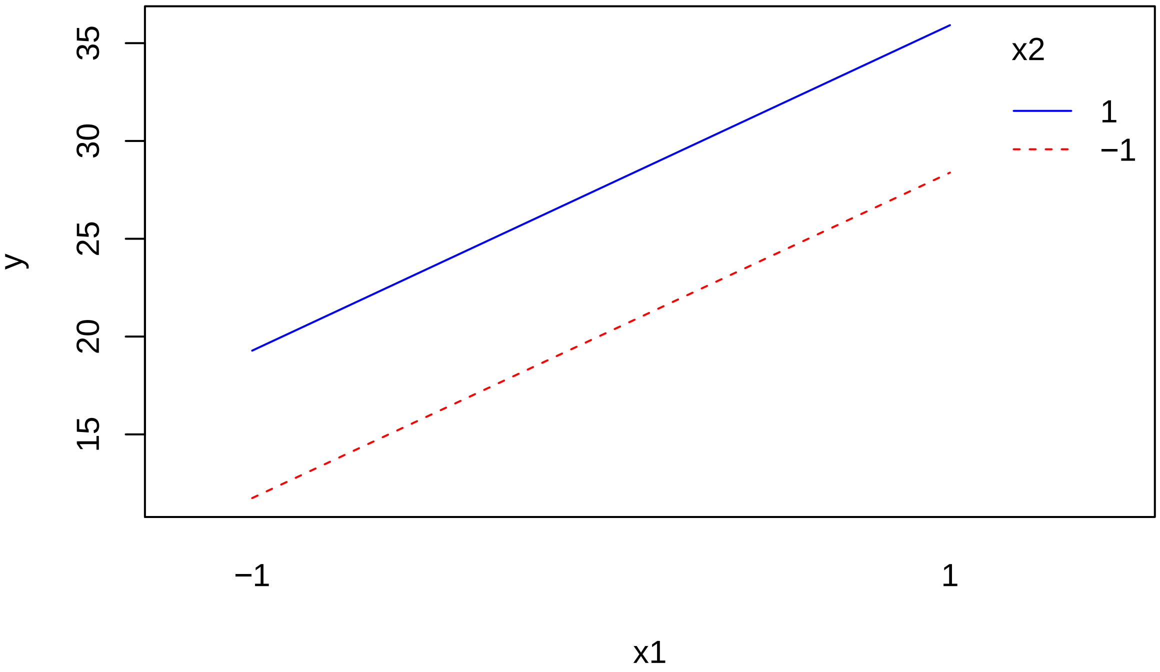
Interaction plot of the interaction model:

```
x.pred = predict(x.mod)
# applying the plot to x1,x2,x.pred same as
# applying the plot to x1,x2,y.
interaction.plot(data$x1,data$x2,x.pred,
                 col=c("red","blue"),
                 trace.label="x2",
                 xlab="x1",
                 ylab="y")
```



Interaction plot of the additive model:

```
add.mod = lm(y~x1+x2,data=data)
add.pred = predict(add.mod)
interaction.plot(data$x1,data$x2,add.pred,
                 col=c("red","blue"),
                 trace.label="x2",
                 xlab="x1",
                 ylab="y")
```



Part (e)

Test for an interaction effect. (Compute the test statistic and p-value.) Provide an interpretation, stated in the context of the problem.

```
anova(add.mod,x.mod)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
13	375.3531	NA	NA	NA	NA
12	71.7225	1	303.6306	50.8009	1.2e-05

$F^* \approx 50$ is quite large ($p = .000$). We find the data to be incompatible with the additive model.

Problem 2

A sample of healthy females is selected to investigate the relationship between age (x) and the level of a steroid (y). Refer to the data from Exercise 8.6.

Part (a)

Fit a quadratic model using coded variables. Compute R^2 for the quadratic model.

```
data = read.table('CH08PR06.txt')
names(data) = c('y','x')
head(data)
```

y	x
27.1	23
22.1	19
21.9	25
10.7	12

y	x
1.4	8
18.8	12

```r
# poly uses coded variables (orthogonal)
#      => z1 = a1 + b1 x
#         z2 = a2 + b2 x + c2 x^2
quad.mod = lm(y ~ poly(x,2),data=data)
coef(quad.mod)
```

```
## (Intercept) poly(x, 2)1 poly(x, 2)2
##    17.64444    28.16524   -15.90551
```

```r
r2 = summary(quad.mod)$r.squared
r2
```

```
## [1] 0.8143372
```

The estimate for the quadratic regression model is

$$E(Y|x) = 17.644 + 28.165x - 15.906x^2,$$

for which

$$R^2 = 0.8143372.$$

Part (b)

Fit a cubic model using coded variables. Compute R^2 for the cubic model.

```r
cubic.mod = lm(y ~ poly(x,3),data=data)
coef(cubic.mod)
```

```
## (Intercept) poly(x, 3)1 poly(x, 3)2 poly(x, 3)3
##   17.644444   28.165236  -15.905513    4.086111
```

```r
r2 = summary(cubic.mod)$r.squared
```

The estimate for the cubic regression model is

$$E(Y|x) = 17.644 + 28.165x - 15.906x^2 + 4.086x^3,$$

which, given the orthogonal design, has the same coefficients as previously for the lower-order terms.

As expected, R^2 has increased,

$$R^2 = 0.8273324,$$

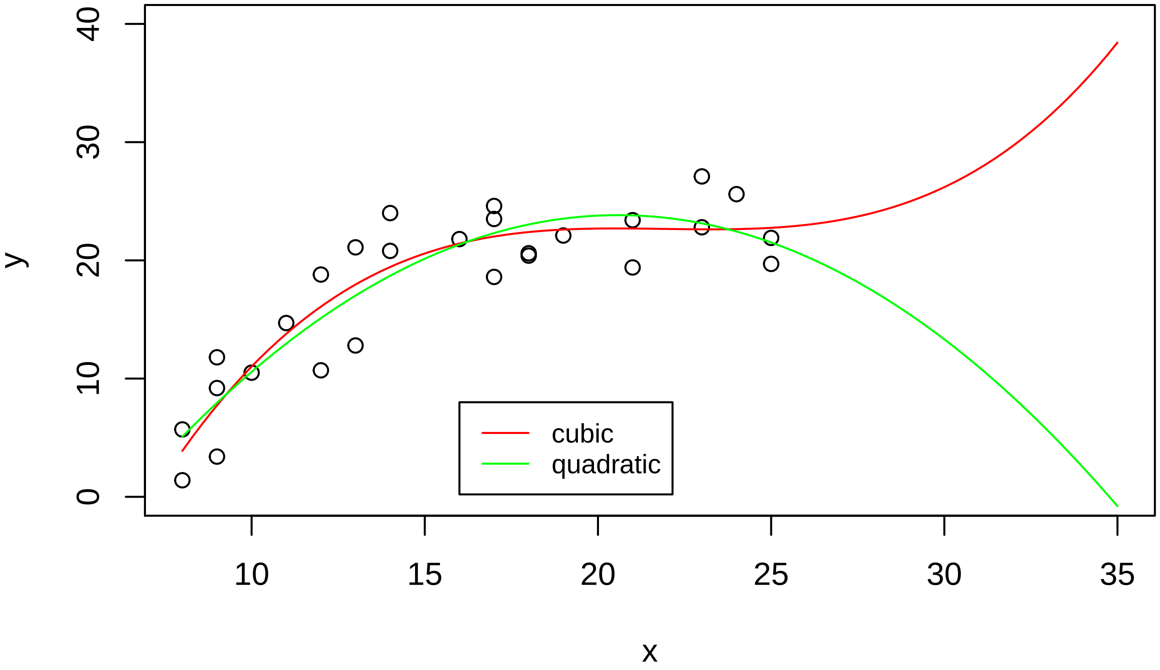
but only slightly.

Part (c)

Create a scatterplot of the data, comparing the fitted values from the quadratic model with the fitted values from the cubic model. Extrapolate the predictions 10 years beyond the largest age in the data set.

```r
newdat = data.frame(x=seq(min(data$x),max(data$x)+10,length.out=100))
newdat$cubic.pred = predict(cubic.mod, newdata = newdat)
newdat$quad.pred = predict(quad.mod, newdata = newdat)
```

```
plot(data$x,data$y,
     xlim=c(min(data$x),max(data$x)+10),
     ylim=c(0,40),
     xlab="x",
     ylab="y")
legend(16, 8, legend=c("cubic","quadratic"),
       col=c("red","green"), lty=1:1, cex=0.8)
points(newdat$x,newdat$cubic.pred,type="l",col="red")
points(newdat$x,newdat$quad.pred,type="l",col="green")
```



Informally, unless we have prior information (e.g., scientific insight), the simpler (quadradic) model seems like a more likely fit to the underlying data generating process.