Regression Analysis - STAT 482 - Probem Set 8

Alex Towell (atowell@siue.edu)

Problem 1

We are interested in modeling the relationship among the predictor variables for the body fat example. Specifically, we wish to model midarm circumference (w) as a function of triceps skinfold thickness (x_1) and thigh circumference (x_2). Refer to the data from Table 7.1. The data for x_1 is listed in the first column, x_2 is listed in the second column, and w is listed in the third column. We are not interested in the body fat measurements, listed in the fourth column, for this problem.

Part (a)

Compute the correlation matrix for w, x_1, x_2.

We drop the last column of data, body fat, from the data set since we are strictly looking at the relations between w (midarm), x_1 (triceps), and x_2 (thigh).

```
data = read.csv('TABLE0701.csv')[,1:3]
head(data)
```

triceps	thigh	midarm
19.5	43.1	29.1
24.7	49.8	28.2
30.7	51.9	37.0
29.8	54.3	31.1
19.1	42.2	30.9
25.6	53.9	23.7

Here is the correlation matrix:

```
data.cor = cor(data)
knitr::kable(round(data.cor,digits=3))
```

	triceps	thigh	midarm
triceps	1.000	0.924	0.458
thigh	0.924	1.000	0.085
midarm	0.458	0.085	1.000

Part (b)

Test for a marginal effect of x_2 on w against a model which includes no other input variables. (Compute the test statistic and p-value.) Provide an interpretation of the result, stated in the context of the problem.

We test for a marginal effect of x_2 on w by comparing the effects model m_2 with the no effects models m_0

where

$$m_0 : w_i = \beta_0 + \epsilon_i$$
$$m_2 : w_i = \beta_0 + \beta_1 x_{2i} + \epsilon_i.$$

We compute the statistics with:

```
names(data) = c("x1","x2","w")
m0 = lm(w~1,  data=data)
m2 = lm(w~x2, data=data)
print(anova(m0,m2))
```

```
## Analysis of Variance Table
##
## Model 1: w ~ 1
## Model 2: w ~ x2
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     19 252.73
## 2     18 250.92  1    1.8117 0.13 0.7227
```

We see that $F_2^* = .130$ with p-value .723.

This is a very large p-value, and so x_2 (thigh) is not adding much explanatory power compared to the model m_0 with no explanatory inputs. In other words, x_2 provides very little predictive power of w (midarm).

Interpretation

The observed data is compatible with the reduced (no effects) model m_0. It is not necessary to add thigh measurement (x_2) to the no effects model for predicting midarm measurement (w).

Part (c)

> Test for a partial effect of x_2 on w against a model which includes x_1. (Compute the test statistic and p-value.) Provide an interpretation of the result, stated in the context of the problem.

We test for a partial effect of x_2 on w given that x_1 is already in the model by comparing models m_1 and m_{12} where

$$m_1 : w_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i,$$
$$m_{12} : w_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{2i} + \epsilon_i.$$

We compute the statistics with:

```
m1 = lm(w~x1, data=data)
m12 = lm(w~x1+x2, data=data)
anova(m1,m12)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
18	199.769500	NA	NA	NA	NA
17	2.416037	1	197.3535	1388.641	0

We see that $F_{2|1}^* = 1388.641$ with a p-value .000.

Interpretation

The observed data is not compatible with the reduced model m_1. We accept the addition of thigh circumference (x_2) as a predictor for midarm circumference (w) to the model which already includes triceps (x_1).

Part (d)

Fit the regression model for w which includes both x_1 and x_2.

```
m12
```

```
##
## Call:
## lm(formula = w ~ x1 + x2, data = data)
##
## Coefficients:
## (Intercept)           x1           x2
##      62.331        1.881       -1.608
```

We estimate that

$$\widehat{w} = 62.331 + 1.881x_1 - 1.608x_2,$$

or using more descriptive names,

$$\widehat{\texttt{midarm}} = 62.331 + 1.881 \times \texttt{triceps} - 1.608 \times \texttt{thigh}.$$

Part (e)

What feature of multidimensional modeling is illustrated in this problem?

Multicollinearity, i.e., highly correlated inputs. Specifically, observe that x_1 and x_2 are strongly positively correlated, $r_{12} = 0.9238425$, but x_1 and x_2 have, respectively, a positive and negative partial effect on w. The combination of these partial effects and the correlation of x_1 and x_2 cancels out their total effect on w.

Deceptively, if we look at the scatterplots of x_1 versus w and x_2 vs w, they seem relatively uncorrelated. Investigating relationships in higher dimensions requires higher level statistical methods, such as regression analysis, rather than two-dimensional methods and graphs.

Additional comments

Much of the data in science, e.g., astronomy, comes from observational studies. Suppose we make an observation of the random sample $\mathcal{D} = \{(X_{i1}, X_{i2}, X_{i3})\}_{i=1}^n$. If we are interested in X_1 and wish to, say, estimate its mean response, then $\mathrm{E}(X_1|X_2, X_3)$ is better than $\mathrm{E}(X_1)$ since $\mathrm{var}(X_1|X_2, X_3) \leq \mathrm{var}(X_1)$.

A controlled experimental design can tell us much more, of course, but we often must make do with what data we already have.

A question one might have is, should we take a sub-sample of $\mathcal{D}$ such that we better approximate an orthogonal design? Honestly, this question seems a bit silly – unless the data is garbage (garbage in, garbage out), we should rarely throw away data. We are generally better off being aware of the issues with multicollinearity and doing our analysis based off the complete sample.

Problem 2

A small scale experiment is conducted to investigate the relationship between crew productivity (y) and crew size (x_1) and bonus pay (x_2). Refer to the data from Table 7.6.

Part (a)

> Provide a definition for an orthogonal design. Discuss an advantage to using an orthogonal design.

A design is orthogonal if $X'X$ is diagonal.

Since $X'X$ is diagonal, the covariance matrix $\text{cov}(b)$ is diagonal, and thus the components of $b = (b_0\ b_1\ \ldots\ b_p)'$ are uncorrelated.

Since b is normally distributed, the mean vector and covariance matrix of any subset of components may be obtained by simply dropping the component indexes to be averaged over. Thus, for orthogonal designs, regression coefficient estimates and variation explained by an input do not depend on which other inputs are included in the model.

Finally, in orthogonal designs, marginal effects and partial effects are the same.

Part (b)

> Fit a multiple regression model using coded inputs. Compute the coefficient estimate b_l, the standard error $\text{se}(b_l)$, the t-statistic, and the p-value, for each of the coded input variables.

```
data = read.table('CH07TA06.txt')
names(data) = c("x1","x2","y")
attach(data)

c1 = 2*(x1-mean(x1))/(range(x1)[2]-range(x1)[1])
c2 = 2*(x2-mean(x2))/(range(x2)[2]-range(x2)[1])

orthog.mod = lm(y~c1+c2)
summary(orthog.mod)
```

```
##
## Call:
## lm(formula = y ~ c1 + c2)
##
## Residuals:
##      1      2      3      4      5      6      7      8
##  1.625 -1.375 -1.625  1.375 -2.125  1.875  0.625 -0.375
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  50.3750     0.6638  75.889 7.53e-09 ***
## c1            5.3750     0.6638   8.097 0.000466 ***
## c2            4.6250     0.6638   6.968 0.000937 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.877 on 5 degrees of freedom
## Multiple R-squared:  0.958,  Adjusted R-squared:  0.9412
## F-statistic: 57.06 on 2 and 5 DF,  p-value: 0.000361
```

Both inputs are statistically significant, and so we select the full additive model.

The computed values for the statistics may be read of the table. For instance, we see that $b_0 = 50.375$, $b_1 = 5.375$, and $b_2 = 4.6250$.