Regression Analysis - STAT 482 - Probem Set 7

Alex Towell (atowell@siue.edu)

The goal is to study the relationship between patient satisfaction (y) and patient's age (x_1), severity of illness (x_2), and anxiety level (x_3). Refer to the data from Exercise 6.15.

> Problem 1
>
> Compute t statistics for testing the effect of each input variable. Explain what type of effect is being tested here.

Computations

```
data = read.table('CH06PR15.txt')
names(data) = c("y", "x1","x2","x3")
add.mod = lm(y ~ x1 + x2 + x3, data=data)
coef(summary(add.mod))
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 158.4912517 18.1258887  8.7439162 5.260955e-11
## x1           -1.1416118  0.2147988 -5.3147960 3.810252e-06
## x2           -0.4420043  0.4919657 -0.8984452 3.740702e-01
## x3          -13.4701632  7.0996608 -1.8972967 6.467813e-02
```

The t statistics are given by

$$\begin{aligned} t_1^* &= -5.315 \\ t_2^* &= -0.898 \\ t_3^* &= -1.897. \end{aligned}$$

Explanation

The statistic t_ℓ^* is testing the partial effect of input x_ℓ, accounting for the effects of all other inputs. (Recall that in the additive model, β_ℓ represents the ℓ-th input effect, with all other inputs held fixed.)

> Problem 2
>
> Compute the F statistic for testing the multiple regression model against a no effects model. Explain what type of effect is being tested here.

Computation

```
null.mod = lm(y ~ 1, data=data)
anova(null.mod,add.mod)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
45	13369.304	NA	NA	NA	NA
42	4248.841	3	9120.464	30.05208	0

We see that $F^* = 30.052$.

Explanation

This is a test for the joint effect of (x_1, x_2, x_3) on response y.

> Probem 3
>
> Compute the sequential sum of squares $\text{SS}_\text{R}(X_1)$, $\text{SS}_\text{R}(X_2|X_1)$, $\text{SS}_\text{R}(X_3|X_1, X_2)$. Explain what each sum of squares represents, stated in the context of the problem.

Computation

```
anova(add.mod)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x1	1	8275.3889	8275.3889	81.802627	0.0000000
x2	1	480.9153	480.9153	4.753871	0.0348861
x3	1	364.1595	364.1595	3.599735	0.0646781
Residuals	42	4248.8407	101.1629	NA	NA

We see that

$$\begin{aligned} \text{SS}_\text{R}(X_1) &= 8275.389, \\ \text{SS}_\text{R}(X_2|X_1) &= 480.915, \\ \text{SS}_\text{R}(X_3|X_1, X_2) &= 364.160. \end{aligned}$$

Explanation

$\text{SS}_\text{R}(X_1)$ measures the variation in response (satisfaction) explained by X_1 (age), not accounting for information on X_2 (severity of illness) and X_3 (anxiety level); $\text{SS}_\text{R}(X_2|X_1)$ measures the variation in satisfaction explained by X_2 beyond that explained by X_1 and not accounting for information in X_3; lastly, $\text{SS}_\text{R}(X_3|X_1, X_2)$ measures the variation in satisfaction explained by X_3 beyond that explained by X_1 and X_2.

Additional remarks

Note that

$$\text{SS}_\text{R}(X_1, X_2, X_3) = \text{SS}_\text{R}(X_1) + \text{SS}_\text{R}(X_2|X_1) + \text{SS}_\text{R}(X_3|X_1, X_2)$$

measures the variation in satisfaction explained by X_1, X_2, and X_3. The additive nature of SS_R is reminiscent of Shannon information H, where if X_1, X_2, X_3 are random variables, then

$$\text{H}(X_1, X_2, X_3) = \text{H}(X_1) + \text{H}(X_2|X_1) + \text{H}(X_3|X_1, X_2).$$

If X_1, X_2, and X_3 are independent, then $\text{H}(X_1, X_2, X_3) = \text{H}(X_1) + \text{H}(X_2) + \text{H}(X_3)$. Likewise, if X_1, X_2, and X_3 explain independent aspects of the variation in the response, I speculate that $\text{SS}_\text{R}(X_1, X_2, X_3) =$

$\text{SS}_\text{R}(X_1) + \text{SS}_\text{R}(X_2) + \text{SS}_\text{R}(X_3)$. Unlike joint probabilities or relative likelihoods, which are multiplicative, information (entropy) and SS_R are additive. This additive property of information is desirable since, for instance, having k times as large of an i.i.d sample for, say, estimating θ, should convey k times as much information about θ (i.e., less variance in an estimator of θ based on the information in such a sample).

That said, it is not clear to me how to fit SS_R into this conceptual framework, if indeed it would even be prudent to do so.

Problem 4

Consider testing for the effect of anxiety level on patient satisfaction.

(a) Test for an effect of X_3 against a model which only includes X_1 (i.e., compute $F^*_{3|1}$ and its p-value).

(b) Test for an effect of X_3 against a model which includes both X_1 and X_2 (i.e., compute $F^*_{3|1,2}$ and its p-value).

(c) Explain the different motivations for the above tests, stated in the context of the problem.

Part (a)

```
m1 = lm(y ~ x1, data=data)
m13 = lm(y ~ x1+x3, data=data)
anova(m1,m13)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
44	5093.915	NA	NA	NA	NA
43	4330.500	1	763.4158	7.58039	0.0086098

We see that $F^*_{3|1} = 7.580$ (p-value = .009). When we already have X_1 in the model, adding X_3 would be an appropriate decision.

Part (b)

```
m12 = lm(y ~ x1+x2, data=data)
m123 = add.mod
anova(m12,m123)
```

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
43	4613.000	NA	NA	NA	NA
42	4248.841	1	364.1595	3.599735	0.0646781

We see that $F^*_{3|1,2} = 3.600$ (p-value = .065). When we already include X_1 and X_2 in the model, the need for X_3 is not as great as before.

Part (c)

In part (a), we are testing for the effect of anxiety level on patient satisfaction after accounting for the effect of patient age.

In part (b), we are testing for the effect of anxiety level on patient satisfaction after accounting for both the effect of patient age and severity of illness.