625.661 Statistical Models and Regression

Test 2 for Modules 5 & 6

H.M. James Hung

1. Prove that in using a regression model analysis to compare the differences of the expected values of the response variable y between the K levels of a categorical regressor x, the sum of squares, SS_T, SS_R, SS_{Res}, will not change regardless of how the K-1 indicators of x are coded (Recall any dummy variable D can be coded in many ways, e.g., D = 0, 1, or, D = -1, 1, or others). [10 points]

State assumptions in each step of your proof.

The key to the proof is to show the equivalence of using a regression model to using the analysis of variance, regardless of how the indicators are coded. This has been shown in page 275 – 280 of the textbook and also in Lecture Mod06B_Indicator Variable.

2. In a study, there are four treatments (labeled as 1, 2, 3, 4) to compare. Assume that there are m subjects per treatment.

(a) Construct an analysis of variance model to compare the four treatments; that is, test whether there is at least one pair of treatments that differ and construct an estimator of every pair of expected treatment difference. [20 points]

$\boldsymbol{y_{1j} = \beta_0 + \beta_1 + \varepsilon_{1j}}$, $\boldsymbol{y_{2j} = \beta_0 + \beta_2 + \varepsilon_{2j}}$, $\boldsymbol{y_{3j} = \beta_0 + \beta_3 + \varepsilon_{3j}}$, $\boldsymbol{y_{4j} = \beta_0 - \beta_1 - \beta_2 - \beta_3 + \varepsilon_{3j}}$.

Thus, $\boldsymbol{\mu_1 = \beta_0 + \beta_1}$, $\boldsymbol{\mu_2 = \beta_0 + \beta_2}$, $\boldsymbol{\mu_3 = \beta_0 + \beta_3}$, $\boldsymbol{\mu_4 = \beta_0 - \beta_1 - \beta_2 - \beta_3}$.

We then have

$\boldsymbol{\mu_1 + \mu_2 + \mu_3 + \mu_4 = 4\beta_0}$, $\boldsymbol{\beta_0 = \frac{\mu_1+\mu_2+\mu_3+\mu_4}{3} \equiv \bar{\mu}}$, $\boldsymbol{\beta_1 = \mu_1 - \bar{\mu}}$, $\boldsymbol{\beta_2 = \mu_2 - \bar{\mu}}$, $\boldsymbol{\beta_3 = \mu_3 - \bar{\mu}}$

Then, μ_i is estimated by the sample mean

$$\bar{y}_{i.} = \frac{1}{m}\sum_{j=1}^{m} y_{ij} \ , \quad i = 1, 2, 3, 4$$

$\bar{\mu}$ **is estimated by** $\bar{y}_{..} = \frac{1}{4}\sum_{i=1}^{4} \bar{y}_{i.}$

Thus, the estimator of the mean difference between treatment h and treatment k is $\bar{y}_{h.} - \bar{y}_{k.}$, where $h = 1, 2, 3, 4$; $k = 1, 2, 3, 4$; $h \neq k$.

We can then test whether there is at least one pair of treatments that Differ by using the analysis of variance table like Table 8.4 on page 276.

(b) Construct a <u>linear regression</u> model such that the regression analysis is equivalent to the analysis of variance in (a). [20 points]

$$y = \begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1m} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2m} \\ y_{31} \\ y_{32} \\ \vdots \\ y_{3m} \\ y_{41} \\ y_{42} \\ \vdots \\ y_{4m} \end{bmatrix}$$

The design matrix X is a $(4m) \times 4$ matrix with

$X_{11} = 1$, $X_{12} = 1$, $X_{13} = 0$, $X_{14} = 0$ **, for each row of treatment 1 group**

$X_{21} = 1$, $X_{22} = 0$, $X_{23} = 1$, $X_{24} = 0$ **, for each row of treatment 2 group**

$X_{31} = 1$, $X_{32} = 0$, $X_{33} = 0$, $X_{34} = 1$ **, for each row of treatment 3 group**

$X_{41} = 1$, $X_{42} = -1$, $X_{43} = -1$, $X_{44} = -1$ **, for each row of treatment 4 group.**

The regression coefficients corresponding to the four columns of X are in the order of $\beta_0, \beta_1, \beta_2, \beta_3$. Thus the regression model in a vector form is $y = XB + \varepsilon$, where $B = (\beta_0, \beta_1, \beta_2, \beta_3)'$ and ε is the random error vector.

$$SS_R(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = \hat{\beta}' X' y$$

$$= (\,\bar{y}_{..} \quad \bar{y}_{1.} - \bar{y}_{..} \quad \bar{y}_{2.} - \bar{y}_{..} \quad \bar{y}_{3.} - \bar{y}_{..}) \begin{pmatrix} y_{..} \\ y_{1.} - y_{4.} \\ y_{2.} - y_{4.} \\ y_{3.} - y_{4.} \end{pmatrix}$$

$$= y_{..}\bar{y}_{..} + (y_{1.} - y_{4.})(\bar{y}_{1.} - \bar{y}_{..}) + (y_{2.} - y_{4.})(\bar{y}_{2.} - \bar{y}_{..}) + (y_{3.} - y_{4.})(\bar{y}_{3.} - \bar{y}_{..})$$

$$= (y_{1.} + y_{2.} + y_{3.} + y_{4.})\bar{y}_{..} + y_{1.}(\bar{y}_{1.} - \bar{y}_{..}) + y_{2.}(\bar{y}_{2.} - \bar{y}_{..}) + y_{3.}(\bar{y}_{3.} - \bar{y}_{..}) \\ - y_{4.}(\bar{y}_{1.} + \bar{y}_{2.} + \bar{y}_{3.} - 3\bar{y}_{..})$$

$$= y_{1.}\bar{y}_{1.} + y_{2.}\bar{y}_{2.} + y_{3.}\bar{y}_{3.} + y_{4.}(4\bar{y}_{..} - \bar{y}_{1.} - \bar{y}_{2.} - \bar{y}_{3.}) = y_{1.}\bar{y}_{1.} + y_{2.}\bar{y}_{2.} + y_{3.}\bar{y}_{3.} + y_{4.}\bar{y}_{4.} \ ,$$

which is exactly the same as the usual sum of squares from the four groups.

Do not use any math/stat software for calculation for Problem 3, except for obtaining the percentile of standard normal, t, chi-square, or F distribution, matrix operations, or basic mathematical calculations.

3. Ten observations on the response variable y associated with two regressor variables x_1 and x_2 are given in the following table. The model fitted to these observations is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \gamma x_{1i} x_{2i} + \varepsilon_i \ , \quad i = 1, \ldots, n \ , \qquad (1)$$

where ε's are identically and independently distributed as a normal random variable with mean zero and a known variance $\sigma^2 = 4$.

Observation #	y	x_1	x_2
1	7	9	1
2	8	6	1
3	5	10	1
4	4	8	1
5	2	5	1
6	10	7	-1
7	9	6	-1
8	10	5	-1
9	8	5	-1
10	8	4	-1

a) Test the null hypothesis "there is no difference between the y-intercept for $x_2 = 1$ and the y-intercept for $x_2 = -1$ and there is no difference between the slope for $x_2 = 1$ and the slope for $x_2 = -1$" at a statistical significance level of 0.05. [20 pts]

For x_2 = 1, the model yields $y = \beta_0 + \beta_1 x_1 + \beta_2 + \gamma x_1 + \varepsilon$;

***y*-intercept = $\beta_0 + \beta_2$, the slope = $\beta_1 + \gamma$.**

For x_2 = -1, the model yields $y = \beta_0 + \beta_1 x_1 - \beta_2 - \gamma x_1 + \varepsilon$;

***y*-intercept = $\beta_0 - \beta_2$, the slope = $\beta_1 - \gamma$.**

The difference in *y*-intercept between x_2 = 1 and x_2 = -1 is $2\beta_2$.

The difference in the slope between x_2 = 1 and x_2 = -1 is 2γ .

Thus, the null hypothesis is H_0: $\beta_2 = 0$, $\gamma = 0$.

$$\mathbf{X'X} = \begin{bmatrix} 10 & 65 & 0 & 11 \\ 65 & 457 & 11 & 155 \\ 0 & 11 & 10 & 65 \\ 11 & 155 & 65 & 457 \end{bmatrix} \qquad \mathbf{X'Y} = \begin{pmatrix} 71 \\ 449 \\ -19 \\ -43 \end{pmatrix}$$

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \gamma \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 4.35 \\ 0.45 \\ -1.54 \\ -0.13 \end{pmatrix} \qquad SS_{res} = 23.3799$$

$$\widehat{Var}\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \gamma \end{pmatrix} = \begin{bmatrix} 9.12 & -1.44 & -2.19 & 0.58 \\ -1.44 & 0.24 & 0.58 & -0.13 \\ -2.19 & 0.58 & 9.12 & -1.44 \\ 0.58 & -0.13 & -1.44 & 0.24 \end{bmatrix}$$

Under H_0: $\beta_2 = 0$, $\gamma = 0$, the model becomes $y = \beta_0 + \beta_1 x_1 + \varepsilon$. This reduced regression model analysis will yield $SS_{Res} = 58.36$.

To test H_0: $\beta_2 = 0$, $\gamma = 0$, we can use *F* test

$$F = \frac{(58.36 - 23.38)/2}{23.38/6} = 4.49$$

The critical value is $F_{0.05;2,6} = 5.14$. Thus, we cannot reject H_0 at $\alpha = 0.05$.

b) Estimate the difference, $E(y \mid x_1 = 5, x_2 = 1) - E(y \mid x_1 = 5, x_2 = -1)$, and calculate its 95% confidence interval. [10 pts]

$$E(y \mid x_1 = 5, x_2 = 1) = \beta_0 + \beta_1 \times 5 + \beta_2 + \gamma \times 5$$
$$E(y \mid x_1 = 5, x_2 = -1) = \beta_0 + \beta_1 \times 5 - \beta_2 - \gamma \times 5$$

$$E(y \mid x_1 = 5, x_2 = 1) - E(y \mid x_1 = 5, x_2 = -1) = 2\beta_2 + 10\gamma$$

Thus, the estimated difference is $2 \times (-1.54) + 10 \times (-0.13) = 4.38$.
The estimated variance of this estimated difference is:

$$(0\ \ 0\ \ 2\ \ 10)\widehat{Var}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \hat{\beta}_2\\ \gamma\end{pmatrix}\begin{pmatrix}0\\0\\2\\10\end{pmatrix} = 4\times 9.12 + 100\times 0.24 -$$

$$2\times 2\times 10\times 1.44 = 2.88$$

Thus, 95% confidence interval of this difference using *t* – distribution with 6 degrees of freedom is $4.38 \mp 2.45 * \sqrt{2.88}$.

c) Predict the difference in y value at $x_1 = 5$ between $x_2 = 1$ and $x_2 = -1$. [10 pts]

The predicted difference in y value at $x_1 = 5$ between $x_2 = 1$ and $x_2 = -1$ is equal to the estimated value in b), that is, it is 4.38.
The estimated variance of the predicted difference is 3.90+2.88 = 6.78
Thus, 95% confidence interval of the predicted difference using *t* – distribution with 6 degrees of freedom is $4.38 \mp 2.45 * \sqrt{6.78}$.

d) Now fit Model (2): $y_i = \beta_0 + \beta_2 x_{2i} + \varepsilon_i$ to the 10 observations.
Calculate the residual for the observation #8 and its variance. [10 pts].

State assumptions in your derivations and calculations in a), b), c), d).

First, $\bar{x}_2 = 0$.

$$Var(\hat{\beta}_0 + \hat{\beta}_2 x_2) = Var(\hat{\beta}_0 + \hat{\beta}_2(x_2 - \bar{x}_2)) = Var(\bar{y}) + Var(\hat{\beta}_2)\ ,$$

because $x_2^2 = 1$ **and** $Cov(\bar{y}, \hat{\beta}_2) = 0$.

Fitting model (2) yields $\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_2\end{pmatrix} = \begin{pmatrix}7.1\\ -1.9\end{pmatrix}$.

For Obs #8, $y = 10$**, the fitted value** $\hat{y} = 7.1 - 1.9\times(1) = 5.2$.

Thus, the residual is $e = 10 - 5.2 = 4.8$.

$$Var(e) = Var(y - \hat{y}) = Var(y) + Var(\hat{y}) - 2Cov(y, \hat{y})$$

$$Cov(y, \hat{y}) = Cov(\hat{y} + e, \hat{y}) = Var(\hat{y}) + 0 = Var(\hat{y})$$

Thus,

$$Var(e) = Var(y - \hat{y}) = Var(y) - Var(\hat{y})$$

Now

$$Var(\hat{y}) = Var\left(\hat{\beta}_0 + \hat{\beta}_2(1)\right) = Var(\bar{y}) + Var(\hat{\beta}_2)\,.$$

$$Var(\bar{y}) = \sigma^2/n$$

$$Var\left(\hat{\beta}_2\right) = \frac{\sigma^2}{\sum(x_{2i}-\bar{x}_2)^2} = \sigma^2/n$$

Thus, $Var(\hat{y}) = Var(\bar{y}) + Var\left(\hat{\beta}_2\right) = 2\sigma^2/n$

$$Var(e) = Var(y - \hat{y}) = Var(y) - Var(\hat{y}) = \sigma^2\left(1 - \frac{2}{n}\right) = 4\left(1 - \frac{2}{10}\right) = 3.2 \quad \text{or} \quad 3.35\left(1 - \frac{2}{10}\right) = 2.7$$

[here 4 is the given variance; 3.35 is estimated σ^2 by mean square error]