STAT 420

Practice Problems 2b

Fall 2015

1. Consider an ANCOVA (analysis of covariance) problem of comparing the means of three populations. Let $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, and $\vec{\mathbf{v}}_3$ be the population indicators.

Sample 1		Sample 2		Sample 3	
y	x	y	x	y	x
35	12	31	7	37	9
34	13	33	11	37	12
35	11	35	13	39	11
35	14	34	13	40	14
36	16	35	16	39	17
38	15	36	15	42	16
39	17	36	16	41	17
40	18	37	17	42	18
41	19	38	18	43	21

Consider the model $\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{v}}_1 + \mu_2 \vec{\mathbf{v}}_2 + \mu_3 \vec{\mathbf{v}}_3 + \beta_4 \vec{\mathbf{x}} + \vec{\mathbf{e}}$.

Test $\mathrm{H}_0: \mu_1 = \mu_2 = \mu_3$ at a 5% level of significance.

```
> sum(lm(y ~ 1)$residuals^2)
[1] 240
> sum(lm(y ~ x + 0)$residuals^2)
[1] 1047.374
> sum(lm(y ~ x)$residuals^2)
[1] 117.3333
> sum(lm(y ~ v1 + v2 + v3 + 0)$residuals^2)
[1] 126
> sum(lm(y ~ v1 + v2 + v3 + x + 0)$residuals^2)
[1] 27.59630
```

2. Consider an ANCOVA (analysis of covariance) problem of comparing the means of four populations.

Sample 1		Sample 2		Sample 3		Sample 4	
x	y	x	y	x	y	x	y
2	17.3	3	17.7	2	16.1	4	17.6
3	21.1	4	18.5	3	20.9	3	9.8
2	19.3	5	18.3	5	20.5	4	18.6
4	23.9	3	14.7	3	18.9	2	13.0
5	21.7	4	20.5	2	13.1	5	18.4

Consider the model $\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{v}}_1 + \mu_2 \vec{\mathbf{v}}_2 + \mu_3 \vec{\mathbf{v}}_3 + \mu_4 \vec{\mathbf{v}}_4 + \beta_5 \vec{\mathbf{x}} + \vec{\boldsymbol{\varepsilon}}$,

where $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$, and $\vec{\mathbf{v}}_4$ are the population indicators, and $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$.

The following results were obtained: $\hat{\mu}_1 = 14.62$ $\hat{\mu}_2 = 10.77$

$\hat{\mu}_3 = 12.24$ $\hat{\mu}_4 = 8.69$ $\hat{\beta}_5 = 1.89$ $\Sigma(y - \hat{y})^2 = 72.644$

$\Sigma(x - \bar{x})^2 = 22.8$ $\Sigma(y - \bar{y})^2 = 213.87$ $\Sigma(x - \bar{x})(y - \bar{y}) = 34.14$

Test $\mathrm{H}_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$ at a 5% level of significance.

3. Consider an ANCOVA (analysis of covariance) problem of comparing five groups with two covariates, $n = 25$. Let $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$, $\vec{\mathbf{v}}_4$, and $\vec{\mathbf{v}}_5$ be the population indicators. Consider the model $\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{v}}_1 + \mu_2 \vec{\mathbf{v}}_2 + \mu_3 \vec{\mathbf{v}}_3 + \mu_4 \vec{\mathbf{v}}_4 + \mu_5 \vec{\mathbf{v}}_5 + \beta_1 \vec{\mathbf{x}}_1 + \beta_2 \vec{\mathbf{x}}_2 + \vec{\boldsymbol{\varepsilon}}$.

```
> sum(lm(y ~ v1 + v2 + v3 + v4 + v5 + x1 + x2 + 0)$residuals^2)
[1] 144
> sum(lm(y ~ v1 + v2 + v3 + v4 + v5 + x1 + 0)$residuals^2)
[1] 164
> sum(lm(y ~ v1 + v2 + v3 + v4 + v5 + x2 + 0)$residuals^2)
[1] 180
> sum(lm(y ~ v1 + v2 + v3 + v4 + v5 + 0)$residuals^2)
[1] 184
> sum(lm(y ~ x1 + x2 + 0)$residuals^2)
[1] 324
> sum(lm(y ~ x1 + x2)$residuals^2)
[1] 240
> sum(lm(y ~ x1 + 0)$residuals^2)
[1] 400
> sum(lm(y ~ x2 + 0)$residuals^2)
[1] 450
> sum(lm(y ~ x1)$residuals^2)
[1] 280
> sum(lm(y ~ x2)$residuals^2)
[1] 300
> sum(lm(y ~ 1)$residuals^2)
[1] 360
```

a) Test $\mathrm{H}_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ at a 5% level of significance.

b) Test $\mathrm{H}_0: \beta_2 = 0$ vs. $\mathrm{H}_1: \beta_2 \neq 0$ at a 10% level of significance.

c) Find the C_p value for the model

$$\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{v}}_1 + \mu_2 \vec{\mathbf{v}}_2 + \mu_3 \vec{\mathbf{v}}_3 + \mu_4 \vec{\mathbf{v}}_4 + \mu_5 \vec{\mathbf{v}}_5 + \vec{\boldsymbol{\varepsilon}}.$$

d) Compute the AIC values for the full model and the model from part (c). Which model is preferred?

Answers:

1. Full model: $\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{v}}_1 + \mu_2 \vec{\mathbf{v}}_2 + \mu_3 \vec{\mathbf{v}}_3 + \beta_4 \vec{\mathbf{x}} + \vec{\mathbf{e}}$

$n = 27$ $\qquad$ $\dim(\mathrm{V}) = 4$

$\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3$

Null model: $\vec{\mathbf{Y}} = \mu \vec{\mathbf{1}} + \beta_4 \vec{\mathbf{x}} + \vec{\mathbf{e}}$

$\dim(\mathrm{V}_0) = 2$

```
> sum(lm(y ~ 1)$residuals^2)
[1] 240
> sum(lm(y ~ x + 0)$residuals^2)
[1] 1047.374
> sum(lm(y ~ x)$residuals^2)                          <- null model
[1] 117.3333
> sum(lm(y ~ v1 + v2 + v3 + 0)$residuals^2)
[1] 126
> sum(lm(y ~ v1 + v2 + v3 + x + 0)$residuals^2)   <- full model
[1] 27.59630
```

	SS	*DF*	*MS*	*F*
Diff.	$\mathrm{RSS}_{\mathrm{null}} - \mathrm{RSS}_{\mathrm{full}}$	$\dim(\mathrm{V}) - \dim(\mathrm{V}_0)$	…	…
Full	$\mathrm{RSS}_{\mathrm{full}}$	$n - \dim(\mathrm{V})$	…	
Null	$\mathrm{RSS}_{\mathrm{null}}$	$n - \dim(\mathrm{V}_0)$		

	SS	*DF*	*MS*	*F*	
Diff.	89.7370	2	44.8685	**37.3954**	← Test Statistic
Full	27.5963	23	1.19984		
Null	117.3333	25			

Critical Value: $\mathrm{F}_{0.05}(2, 23) = \mathbf{3.42}$. $\qquad$ Decision: **Reject $\mathbf{H}_0$**.

2. Full model: $\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{v}}_1 + \mu_2 \vec{\mathbf{v}}_2 + \mu_3 \vec{\mathbf{v}}_3 + \mu_4 \vec{\mathbf{v}}_4 + \beta_5 \vec{\mathbf{x}} + \vec{\boldsymbol{\varepsilon}}$.

Since $\vec{\mathbf{v}}_1 + \vec{\mathbf{v}}_2 + \vec{\mathbf{v}}_3 + \vec{\mathbf{v}}_4 = \vec{\mathbf{1}}$, under $\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 \ (= \mu)$ the full model becomes

Null model: $\vec{\mathbf{Y}} = \mu \vec{\mathbf{1}} + \beta \vec{\mathbf{x}} + \vec{\boldsymbol{\varepsilon}}$ – simple linear regression.

$\mathrm{V} = \{ a_1 \mathbf{v_1} + a_2 \mathbf{v_2} + a_3 \mathbf{v_3} + a_4 \mathbf{v_4} + a_5 \mathbf{x}, \quad a_1, a_2, a_3, a_4, a_5 \in \mathbf{R} \}$,

$\dim(\mathrm{V}) = 5.$

$\mathrm{V}_0 = \{ a_0 \mathbf{1} + a_1 \mathbf{x}, \quad a_0, a_1 \in \mathbf{R} \}$, $\dim(\mathrm{V}_0) = 2.$

Numerator d.f. $= \dim(\mathrm{V}) - \dim(\mathrm{V}_0) = 5 - 2 = \mathbf{3}$.

Denominator d.f. $= n - \dim(\mathrm{V}) = 20 - 5 = \mathbf{15}$.

	SS	*DF*	*MS*	*F*
Diff.	$\mathrm{SSResid}_{\mathrm{null}} - \mathrm{SSResid}_{\mathrm{full}}$	$\dim(\mathrm{V}) - \dim(\mathrm{V}_0)$	…	…
Full	$\mathrm{SSResid}_{\mathrm{full}}$	$n - \dim(\mathrm{V})$	…	
Null	$\mathrm{SSResid}_{\mathrm{null}}$	$n - \dim(\mathrm{V}_0)$		

$\mathrm{SSResid}_{\mathrm{full}} = 72.644.$

$\mathrm{SSResid}_{\mathrm{null}}$:

$$\hat{\beta} = \frac{\mathrm{SXY}}{\mathrm{SXX}} = \frac{34.14}{22.8}$$

$$\mathrm{SSRegr}_{\mathrm{null}} = \hat{\beta}^2 \, \mathrm{SXX} = \left(\frac{34.14}{22.8} \right)^2 \cdot 22.8 \approx 51.12$$

$$\mathrm{SSResid}_{\mathrm{null}} = \mathrm{SYY} - \mathrm{SSRegr}_{\mathrm{null}} \approx 213.87 - 51.12 = 162.75.$$

	SS	*DF*	*MS*	*F*	
Diff.	90.106	3	30.035	**6.2**	← Test Statistic
Full	72.644	15	4.843		
Null	162.75	18			

Critical Value: $F_{0.05}(3, 15) = \mathbf{3.29}$.

Decision: **Reject $\mathbf{H_0}$**.

3. Consider an ANCOVA (analysis of covariance) problem of comparing five groups with two covariates, $n = 25$. Let $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, $\vec{\mathbf{v}}_3$, $\vec{\mathbf{v}}_4$, and $\vec{\mathbf{v}}_5$ be the population indicators.

Consider the model $\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{v}}_1 + \mu_2 \vec{\mathbf{v}}_2 + \mu_3 \vec{\mathbf{v}}_3 + \mu_4 \vec{\mathbf{v}}_4 + \mu_5 \vec{\mathbf{v}}_5 + \beta_1 \vec{\mathbf{x}}_1 + \beta_2 \vec{\mathbf{x}}_2 + \vec{\boldsymbol{\varepsilon}}$.

```
> sum(lm(y ~ v1 + v2 + v3 + v4 + v5 + x1 + x2 + 0)$residuals^2)
[1] 144
> sum(lm(y ~ v1 + v2 + v3 + v4 + v5 + x1 + 0)$residuals^2)
[1] 164
> sum(lm(y ~ v1 + v2 + v3 + v4 + v5 + x2 + 0)$residuals^2)
[1] 180
> sum(lm(y ~ v1 + v2 + v3 + v4 + v5 + 0)$residuals^2)
[1] 184
> sum(lm(y ~ x1 + x2 + 0)$residuals^2)
[1] 324
> sum(lm(y ~ x1 + x2)$residuals^2)
[1] 240
> sum(lm(y ~ x1 + 0)$residuals^2)
[1] 400
> sum(lm(y ~ x2 + 0)$residuals^2)
[1] 450
> sum(lm(y ~ x1)$residuals^2)
[1] 280
> sum(lm(y ~ x2)$residuals^2)
[1] 300
> sum(lm(y ~ 1)$residuals^2)
[1] 360
```

a) Test $\mathrm{H}_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ at a 5% level of significance.

Full model: $\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{v}}_1 + \mu_2 \vec{\mathbf{v}}_2 + \mu_3 \vec{\mathbf{v}}_3 + \mu_4 \vec{\mathbf{v}}_4 + \mu_5 \vec{\mathbf{v}}_5 + \beta_1 \vec{\mathbf{x}}_1 + \beta_2 \vec{\mathbf{x}}_2 + \vec{\boldsymbol{\varepsilon}}$.

$\dim(\mathrm{V}) = 7.$ $\quad$ $\mathrm{SSResid}_{\mathrm{full}} = 144.$

Since $\vec{\mathbf{v}}_1 + \vec{\mathbf{v}}_2 + \vec{\mathbf{v}}_3 + \vec{\mathbf{v}}_4 + \vec{\mathbf{v}}_5 = \vec{\mathbf{1}}$,

Null model: $\vec{\mathbf{Y}} = \mu \vec{\mathbf{1}} + \beta_1 \vec{\mathbf{x}}_1 + \beta_2 \vec{\mathbf{x}}_2 + \vec{\boldsymbol{\varepsilon}}$.

$\dim(\mathrm{V}_0) = 3.$ $\quad$ $\mathrm{SSResid}_{\mathrm{full}} = 240.$

	SS	DF	MS	F
Diff.	$\text{SSResid}_{\text{null}} - \text{SSResid}_{\text{full}}$	$\dim(\text{V}) - \dim(\text{V}_0)$	…	…
Full	$\text{SSResid}_{\text{full}}$	$n - \dim(\text{V})$	…	
Null	$\text{SSResid}_{\text{null}}$	$n - \dim(\text{V}_0)$		

	SS	DF	MS	F	
Diff.	96	4	24	**3**	← Test Statistic
Full	144	18	8		
Null	240	22			

Critical Value: $\text{F}_{0.05}(4, 18) = \mathbf{2.93}$. Decision: **Reject $\mathbf{H_0}$.**

b) Test $\text{H}_0 : \beta_2 = 0$ vs. $\text{H}_1 : \beta_2 \neq 0$ at a 10% level of significance.

Full model: $\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{v}}_1 + \mu_2 \vec{\mathbf{v}}_2 + \mu_3 \vec{\mathbf{v}}_3 + \mu_4 \vec{\mathbf{v}}_4 + \mu_5 \vec{\mathbf{v}}_5 + \beta_1 \vec{\mathbf{x}}_1 + \beta_2 \vec{\mathbf{x}}_2 + \vec{\boldsymbol{\varepsilon}}$.

$\dim(\text{V}) = 7$. $\text{SSResid}_{\text{full}} = 144$.

Null model: $\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{v}}_1 + \mu_2 \vec{\mathbf{v}}_2 + \mu_3 \vec{\mathbf{v}}_3 + \mu_4 \vec{\mathbf{v}}_4 + \mu_5 \vec{\mathbf{v}}_5 + \beta_1 \vec{\mathbf{x}}_1 + \vec{\boldsymbol{\varepsilon}}$.

$\dim(\text{V}_0) = 6$. $\text{SSResid}_{\text{full}} = 164$.

	SS	DF	MS	F	
Diff.	20	1	20	**2.5**	← Test Statistic
Full	144	18	8		
Null	164	19			

Critical Value: $\text{F}_{0.10}(1, 18) = \mathbf{3.01}$. Decision: **Do NOT Reject $\mathbf{H_0}$.**

c) Find the C_p value for the model

$$\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{v}}_1 + \mu_2 \vec{\mathbf{v}}_2 + \mu_3 \vec{\mathbf{v}}_3 + \mu_4 \vec{\mathbf{v}}_4 + \mu_5 \vec{\mathbf{v}}_5 + \vec{\boldsymbol{\varepsilon}} .$$

$$C_p = \frac{\text{SSResid}_{\text{Null}}}{\text{MSResid}_{\text{Full}}} - n + 2 \cdot (\#\text{ of parameters in Null}) = \frac{184}{8} - 25 + 2 \cdot (5) = \mathbf{8}.$$

[Want models with C_p close to or less than $(\#\text{ of parameters in Null})$.]

d) Compute the AIC values for the full model and the model from part (c). Which model is preferred?

Full: $\text{AIC} = 25 \ln\left({}^{144}/_{25} \right) + 2 \times 7 \approx \mathbf{57.77}$.

Part (c): $\text{AIC} = 25 \ln\left({}^{184}/_{25} \right) + 2 \times 5 \approx \mathbf{59.90}$.

OR

Full: $\text{AIC} = 25 + 25 \ln(2\pi) + 25 \ln\left({}^{144}/_{25} \right) + 2 \times 7 \approx \mathbf{111.39}$.

Part (c): $\text{AIC} = 25 + 25 \ln(2\pi) + 25 \ln\left({}^{184}/_{25} \right) + 2 \times 5 \approx \mathbf{113.52}$.

The **full mode**l is preferred since the full model has lower AIC value.