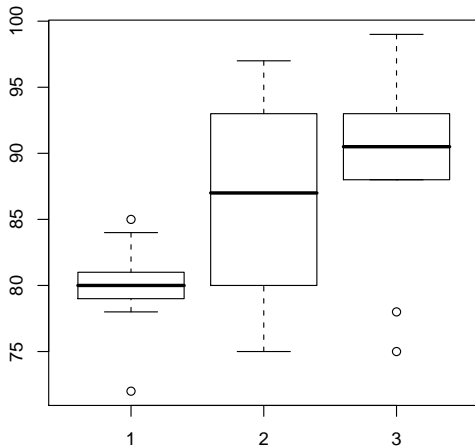


# Linear Models: R Examples — The less than full rank model: inference

# One-factor tests

For one-factor tests, we keep the running example of mathematics marks.



$$H_0 : \tau_1 = \tau_2 = \tau_3$$

```
> C <- matrix(c(0, 1, -1, 0, 0, 0, 1, -1), 2, 4, byrow = TRUE)
> C
```

```
      [,1] [,2] [,3] [,4]
[1,]    0    1  -1    0
[2,]    0    0    1  -1
```

```
> C %*% XtXc %*% t(X) %*% X
```

```
      [,1] [,2] [,3] [,4]
[1,]    0    1  -1    0
[2,]    0    0    1  -1
```

$$H_0 : \tau_1 = \tau_2 = \tau_3$$

```
> SS <- t(C %*% b) %*% inv(C %*% XtXc %*% t(C)) %*% C %*% b
> SS
```

```
      [,1]
[1,] 474.0667
```

```
> Fstat <- (SS/2)/s2
> Fstat
```

```
      [,1]
[1,] 5.624802
```

```
> pf(Fstat, 2, n - r, lower.tail = FALSE)
```

```
      [,1]
[1,] 0.009077098
```

```
> options(contrasts = c("contr.treatment", "contr.poly"))
> model <- lm(maths.y ~ class.f, data = maths)
> summary(model)
```

Call:

```
lm(formula = maths.y ~ class.f, data = maths)
```

Residuals:

Min	1Q	Median	3Q	Max
-14.40	-1.80	0.85	3.60	10.50

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	79.900	2.053	38.922	< 2e-16 ***
class.f2	6.600	2.903	2.273	0.03117 *
class.f3	9.500	2.903	3.272	0.00292 **

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.492 on 27 degrees of freedom

Multiple R-squared: 0.2941, Adjusted R-squared: 0.2418

F-statistic: 5.625 on 2 and 27 DF, p-value: 0.009077

$$H_0 : \tau_1 = \tau_2 = \tau_3 \quad (R)$$

```
> anova(model)
```

```
Analysis of Variance Table
```

```
Response: maths.y
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
class.f	2	474.07	237.03	5.6248	0.009077 **
Residuals	27	1137.80	42.14		

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$H_0 : \tau_1 = \tau_2 = \tau_3 \quad (R)$$

```
> basemodel <- lm(maths.y ~ 1, data = maths)
> anova(basemodel, model)
```

Analysis of Variance Table

Model 1: maths.y ~ 1

Model 2: maths.y ~ class.f

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	29	1611.87				
2	27	1137.80	2	474.07	5.6248	0.009077 **

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

$$H_0 : 2\tau_1 = \tau_2 + \tau_3$$

```
> C <- matrix(c(0, 2, -1, -1), 1, 4)
> SS <- t(C %*% b) %*% inv(C %*% XtXc %*% t(C)) %*% C %*% b
> SS
```

```
      [,1]
[1,] 432.0167
```

```
> Fstat <- (SS/1)/s2
> Fstat
```

```
      [,1]
[1,] 10.25176
```

```
> pf(Fstat, 1, n - r, lower.tail = FALSE)
```

```
      [,1]
[1,] 0.00348348
```

$$H_0 : 2\tau_1 = \tau_2 + \tau_3 \quad (\text{R})$$

The hypothesis is equivalent to  $-(\mu_2 - \mu_1) - (\mu_3 - \mu_1) = 0$ .

```
> linear.hypothesis(model, c(0, -1, -1), 0)
```

Linear hypothesis test

Hypothesis:

```
-class.f2 - class.f3 = 0
```

```
Model 1: maths.y ~ class.f
```

```
Model 2: restricted model
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	27	1137.80				
2	28	1569.82	-1	-432.02	10.252	0.003483 **

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Two-factor models

We look at the effect of pre-chamber volume ratio and injection timing on the emission of noxious gas from an engine. The factors have 3 levels each.

```
> str(engine)
```

```
'data.frame': 18 obs. of 3 variables:
```

```
$ gas : num 6.27 8.08 7.34 5.43 8.04 7.87 6.94 7.48 8.61 6.5
```

```
$ volume: Factor w/ 3 levels "low","medium",...: 1 2 3 1 2 3 1 2
```

```
$ time : Factor w/ 3 levels "short","medium",...: 1 1 1 1 1 1 2
```

```
> means
```

```
      [,1] [,2] [,3]  
[1,] 5.850 6.725 7.135  
[2,] 8.060 7.500 8.810  
[3,] 7.605 8.465 9.045
```

```
> model <- lm(gas ~ time + volume, data = engine)
> summary(model)
```

Call:

```
lm(formula = gas ~ time + volume, data = engine)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.62333	-0.15000	0.03583	0.21375	0.49500

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	6.0533	0.2109	28.703	3.83e-13	***
timemedium	0.3917	0.2310	1.695	0.113817	
timelong	1.1583	0.2310	5.014	0.000237	***
volumemedium	1.5533	0.2310	6.724	1.42e-05	***
volumehigh	1.8017	0.2310	7.798	2.95e-06	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4002 on 13 degrees of freedom

Multiple R-squared: 0.8823, Adjusted R-squared: 0.8461

F-statistic: 24.27 on 4 and 13 DF, p-value: 6.126e-06

# Testing differences among populations

```
> anova(model)
```

```
Analysis of Variance Table
```

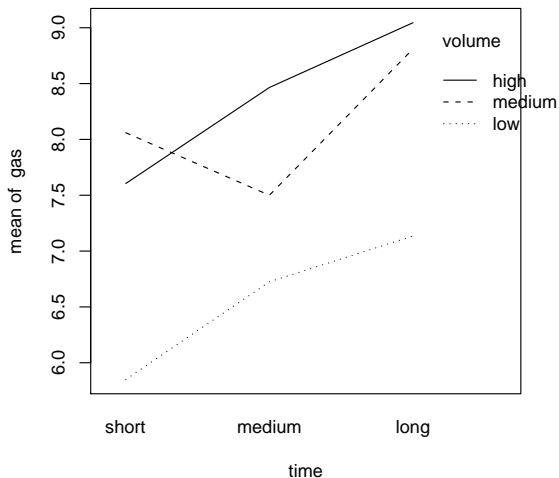
```
Response: gas
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
time	2	4.1658	2.0829	13.008	0.0007898 ***
volume	2	11.4410	5.7205	35.726	5.22e-06 ***
Residuals	13	2.0816	0.1601		

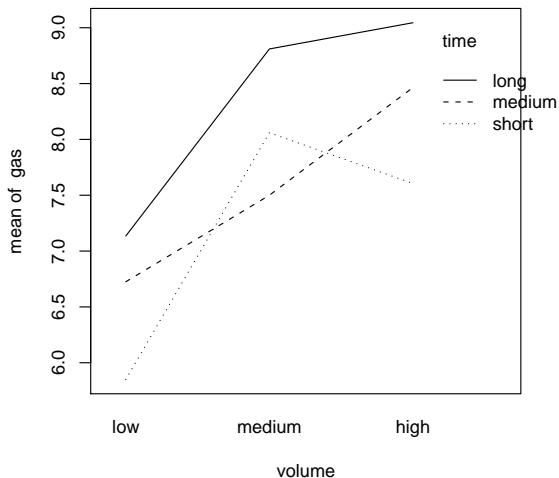
```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> with(engine, interaction.plot(time, volume, gas))
```



```
> with(engine, interaction.plot(volume, time, gas))
```



```
> imodel <- lm(gas ~ time * volume, data = engine)
> summary(imodel)
```

Call:

```
lm(formula = gas ~ time * volume, data = engine)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-4.200e-01	-1.300e-01	6.939e-18	1.300e-01	4.200e-01

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	5.8500	0.1967	29.745	2.68e-10	***
timemedium	0.8750	0.2781	3.146	0.011815	*
timelong	1.2850	0.2781	4.620	0.001254	**
volumemedium	2.2100	0.2781	7.946	2.34e-05	***
volumehigh	1.7550	0.2781	6.310	0.000139	***
timemedium:volumemedium	-1.4350	0.3933	-3.648	0.005333	**
timelong:volumemedium	-0.5350	0.3933	-1.360	0.206882	
timemedium:volumehigh	-0.0150	0.3933	-0.038	0.970413	
timelong:volumehigh	0.1550	0.3933	0.394	0.702715	

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# Testing for interaction

```
> anova(imodel)
```

```
Analysis of Variance Table
```

```
Response: gas
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
time	2	4.1658	2.0829	26.9246	0.0001591	***
volume	2	11.4410	5.7205	73.9456	2.594e-06	***
time:volume	4	1.3853	0.3463	4.4768	0.0289181	*
Residuals	9	0.6963	0.0774			

```
---
```

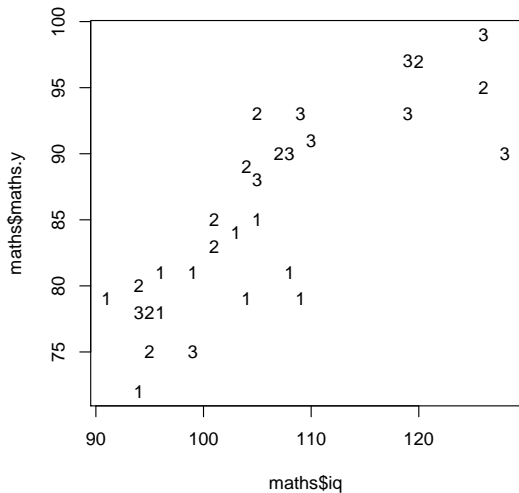
```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## An ANCOVA example

The maths dataset also has another component - the IQ of the student.

```
> str(maths)
```

```
'data.frame': 30 obs. of 5 variables:  
 $ X      : int  1 2 3 4 5 6 7 8 9 10 ...  
 $ maths.y: int  81 84 81 79 78 79 81 85 72 79 ...  
 $ iq     : int  99 103 108 109 96 104 96 105 94 91 ...  
 $ class  : int  1 1 1 1 1 1 1 1 1 1 ...  
 $ class.f: Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1
```



```
> model <- lm(maths.y ~ class.f + iq + class.f:iq, data = maths)
> summary(model)
```

Call:

```
lm(formula = maths.y ~ class.f + iq + class.f:iq, data = maths)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.2507	-1.8312	0.9807	2.4711	6.3765

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	52.7577	21.9941	2.399	0.0246 *
class.f2	-30.9642	25.7058	-1.205	0.2401
class.f3	-24.0093	25.8357	-0.929	0.3620
iq	0.2701	0.2185	1.236	0.2284
class.f2:iq	0.3474	0.2524	1.376	0.1815
class.f3:iq	0.2729	0.2497	1.093	0.2852

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.043 on 24 degrees of freedom

Multiple R-squared: 0.7566, Adjusted R-squared: 0.7059

# Testing for interaction

```
> anova(model)
```

Analysis of Variance Table

Response: maths.y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
class.f	2	474.07	237.03	14.499	7.437e-05	***
iq	1	714.38	714.38	43.697	7.749e-07	***
class.f:iq	2	31.06	15.53	0.950	0.4008	
Residuals	24	392.36	16.35			

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Interaction is not significant, so we remove the interaction term and fit an additive model.

```
> amodel <- lm(maths.y ~ class.f + iq, data = maths)
> summary(amodel)
```

Call:

```
lm(formula = maths.y ~ class.f + iq, data = maths)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.137	-2.842	1.221	2.662	6.393

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	26.02809	8.23338	3.161	0.00396	**
class.f2	4.29503	1.83799	2.337	0.02743	*
class.f3	3.49636	2.01959	1.731	0.09526	.
iq	0.53604	0.08093	6.623	5.03e-07	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.036 on 26 degrees of freedom

Multiple R-squared: 0.7373, Adjusted R-squared: 0.707

F-statistic: 24.33 on 3 and 26 DF, p-value: 1.032e-07

```
> basemodel <- lm(maths.y ~ class.f, data = maths)
> anova(basemodel, amodel)
```

### Analysis of Variance Table

```
Model 1: maths.y ~ class.f
```

```
Model 2: maths.y ~ class.f + iq
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	27	1137.80				
2	26	423.42	1	714.38	43.866	5.032e-07 ***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Clearly IQ is significant.

```
> basemodel <- lm(maths.y ~ iq, data = maths)
> anova(basemodel, amodel)
```

### Analysis of Variance Table

Model 1: maths.y ~ iq

Model 2: maths.y ~ class.f + iq

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	28	518.13				
2	26	423.42	2	94.71	2.9077	0.0725 .

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The class is not very significant. However, since it is so close, we will retain it.

