STAT 420

1. $J = 3$ different brands of corn were planted on 10 plots, $n_1 = 3$, $n_2 = 4$, $n_3 = 3$.
The crop yields y (in bushels per acre) were as follows:

	Brand 1	Brand 2	Brand 3
y	104, 109, 108	104, 104, 110, 106	110, 109, 114

Consider the model

$$\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{v}}_1 + \mu_2 \vec{\mathbf{v}}_2 + \mu_3 \vec{\mathbf{v}}_3 + \vec{\varepsilon}$$

with the usual assumptions on $\vec{\varepsilon}$, where $\vec{\mathbf{v}}_1 = (1, 1, 1,\ 0, 0, 0, 0,\ 0, 0, 0)^T$,
$\vec{\mathbf{v}}_2 = (0, 0, 0,\ 1, 1, 1, 1,\ 0, 0, 0)^T$, $\vec{\mathbf{v}}_3 = (0, 0, 0,\ 0, 0, 0, 0,\ 1, 1, 1)^T$.

Use a 10% significance level to test H_0: $\mu_1 = \mu_2 = \mu_3$. (ANOVA)

Overall:

$$\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{v}}_1 + \mu_2 \vec{\mathbf{v}}_2 + \mu_3 \vec{\mathbf{v}}_3 + \vec{\varepsilon}$$

Under H_0:

$$\vec{\mathbf{Y}} = \mu \vec{\mathbf{1}} + \vec{\varepsilon}$$

```
> y = c(104,109,108, 104,104,110,106, 110,109,114)
> v1 = c(1,1,1, 0,0,0,0, 0,0,0)
> v2 = c(0,0,0, 1,1,1,1, 0,0,0)
> v3 = c(0,0,0, 0,0,0,0, 1,1,1)
>
> fit = lm(y ~ v1 + v2 + v3 + 0)
> summary(fit)

Call:
lm(formula = y ~ v1 + v2 + v3 + 0)

Residuals:
   Min      1Q Median      3Q     Max
 -3.00   -2.00  -0.50    1.75    4.00
```

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
v1  107.000      1.574   68.00 3.91e-11 ***
v2  106.000      1.363   77.78 1.53e-11 ***
v3  111.000      1.574   70.54 3.03e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.726 on 7 degrees of freedom
Multiple R-Squared: 0.9996,     Adjusted R-squared: 0.9994
F-statistic:  5217 on 3 and 7 DF,  p-value: 4.399e-12

> anova(lm(y ~ 1),fit)
Analysis of Variance Table

Model 1: y ~ 1
Model 2: y ~ v1 + v2 + v3 + 0
  Res.Df  RSS Df Sum of Sq      F Pr(>F)
1      9 97.6
2      7 52.0  2      45.6 3.0692 0.1104
```

OR

```
> y = c(104,109,108, 104,104,110,106, 110,109,114)
> brand = c(1,1,1, 2,2,2,2, 3,3,3)
>
> fit0 = glm(y ~ factor(brand))
> summary(aov(fit0))
              Df Sum Sq Mean Sq F value Pr(>F)
factor(brand)  2 45.600  22.800  3.0692 0.1104
Residuals      7 52.000   7.429
```

$F = 3.0692 < F_{0.10}(2, 7) = 3.26.$

P-value $= 0.1104 > \alpha = 0.10.$

Do NOT Reject $H_0: \mu_1 = \mu_2 = \mu_3$ at $\alpha = 0.10$.

2. $J = 3$ different brands of corn were planted on 10 plots, $n_1 = 3$, $n_2 = 4$, $n_3 = 3$. The crop yields y (in bushels per acre) and the amounts x of fertilizer (a blend of nitrogen, phosphate, and potash) used (in pounds per acre) were as follows:

	Brand 1	Brand 2	Brand 3
y	104, 109, 108	104, 104, 110, 106	110, 109, 114
x	10, 20, 30	20, 10, 40, 30	20, 30, 40

Consider the model

$$\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{v}}_1 + \mu_2 \vec{\mathbf{v}}_2 + \mu_3 \vec{\mathbf{v}}_3 + \beta_4 \vec{\mathbf{x}} + \vec{\varepsilon}$$

with the usual assumptions on $\vec{\varepsilon}$, where $\vec{\mathbf{v}}_1 = (1, 1, 1,\ 0, 0, 0, 0,\ 0, 0, 0)^T$, $\vec{\mathbf{v}}_2 = (0, 0, 0,\ 1, 1, 1, 1,\ 0, 0, 0)^T$, $\vec{\mathbf{v}}_3 = (0, 0, 0,\ 0, 0, 0, 0,\ 1, 1, 1)^T$.

Use a 10% significance level to test $H_0 :\ \mu_1 = \mu_2 = \mu_3$. (ANCOVA)

Overall:

$$\vec{\mathbf{Y}} = \mu_1 \vec{\mathbf{v}}_1 + \mu_2 \vec{\mathbf{v}}_2 + \mu_3 \vec{\mathbf{v}}_3 + \beta_4 \vec{\mathbf{x}} + \vec{\varepsilon}$$

Under H_0:

$$\vec{\mathbf{Y}} = \mu \vec{\mathbf{1}} + \beta_4 \vec{\mathbf{x}} + \vec{\varepsilon}$$

```
> y = c(104,109,108, 104,104,110,106, 110,109,114)
> v1 = c(1,1,1, 0,0,0,0, 0,0,0)
> v2 = c(0,0,0, 1,1,1,1, 0,0,0)
> v3 = c(0,0,0, 0,0,0,0, 1,1,1)
> x = c(10,20,30, 20,10,40,30, 20,30,40)
>
> fit1 = lm(y ~ v1 + v2 + v3 + x + 0)
> summary(fit1)

Call:
lm(formula = y ~ v1 + v2 + v3 + x + 0)

Residuals:
       Min         1Q     Median         3Q        Max
-2.000e+00 -1.000e+00  9.992e-15  1.000e+00  2.000e+00
```

```
Coefficients:
    Estimate Std. Error t value Pr(>|t|)
v1 103.00000    1.44016  71.520 5.03e-10 ***
v2 101.00000    1.58698  63.643 1.01e-09 ***
v3 105.00000    1.88562  55.685 2.25e-09 ***
x    0.20000    0.05443   3.674   0.0104 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.633 on 6 degrees of freedom
Multiple R-Squared: 0.9999,     Adjusted R-squared: 0.9998
F-statistic: 1.09e+04 on 4 and 6 DF,  p-value: 1.041e-11

> fit2 = lm(y ~ x)
> summary(fit2)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9429 -1.1571 -0.3714  1.7714  3.3429

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 102.08571    1.92732  52.968 1.79e-11 ***
x             0.22857    0.07133   3.204   0.0125 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.311 on 8 degrees of freedom
Multiple R-Squared: 0.5621,     Adjusted R-squared: 0.5073
F-statistic: 10.27 on 1 and 8 DF,  p-value: 0.01253

> anova(fit2,fit1)
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ v1 + v2 + v3 + x + 0
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1      8 42.743
2      6 16.000  2    26.743 5.0143 0.05245 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$F = 5.0143 > F_{0.10}(2, 6) = 3.46.$ P-value $= 0.05245 < \alpha = 0.10.$

Reject $H_0: \mu_1 = \mu_2 = \mu_3$ at $\alpha = 0.10$.

OR

```
> fit3 = glm(y ~ x + factor(brand))
> summary(aov(fit3))
              Df Sum Sq Mean Sq F value   Pr(>F)
x              1 54.857  54.857 20.5714 0.003952 **
factor(brand)  2 26.743  13.371  5.0143 0.052453 .
Residuals      6 16.000   2.667
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

STAT 420

3. A company wishes to study the effects of three different types of promotion on sales of its cookies. The three promotions were:

Treatment 1 – Sampling of product by customers in store and regular shelf space

Treatment 2 – Special display shelves at ends of aisle in addition to regular shelf space

Treatment 3 – Additional shelf space in regular location

Fifteen stores were selected as the experimental units. Each store was randomly assigned one of the promotion types, with five stores assigned to each type of promotion. Other relevant conditions under the control of the company, such as price and advertising, were kept the same for all stores in the experiment. Data on the number of cases of the product sold during the promotional period, denoted by Y, are presented in the table below, as are also data on the sales of the product in the preceding period, denoted by X. Sales of the preceding period are to be used as the covariate variable.

Experimental Unit	Treatment 1		Treatment 2		Treatment 3	
	Y	X	Y	X	Y	X
1	38	21	43	34	24	23
2	39	26	38	26	32	29
3	36	22	38	29	31	30
4	45	28	27	18	21	16
5	33	19	34	25	28	29

Test for treatment effects (test whether or not the three promotions differ in effectiveness):

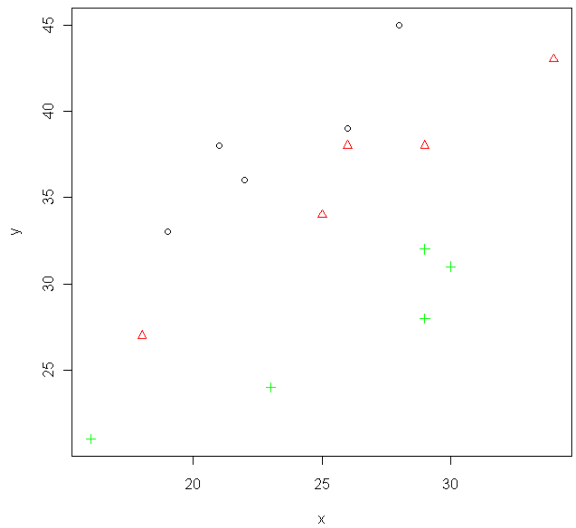
i. Specify the full model.

ii. Specify the null hypothesis H_0 in the notations of your full model.

iii. Specify the model under the null hypothesis H_0.

iv. Conduct the *F* test at significance level $\alpha = 0.05$. (Show the calculations leading to your conclusion in the form of an ANOVA table.) State your decision/conclusion.

```
> Cookies = read.table(" ... /Cookies.csv", sep=",", header=T)
> attach(Cookies)
> Cookies
    y  x v2 v3
1  38 21  0  0
2  39 26  0  0
3  36 22  0  0
4  45 28  0  0
5  33 19  0  0
6  43 34  1  0
7  38 26  1  0
8  38 29  1  0
9  27 18  1  0
10 34 25  1  0
11 24 23  0  1
12 32 29  0  1
13 31 30  0  1
14 21 16  0  1
15 28 29  0  1
> plot(x,y,col=1+v2+2*v3,pch=1+v2+2*v3)
>
> fit = lm(y ~ x + v2 + v3)
> summary(fit)
```




```
Call:
lm(formula = y ~ x + v2 + v3)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4348 -1.2739 -0.3363  1.6710  2.4869

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.3534     2.5230   6.878 2.66e-05 ***
x             0.8986     0.1026   8.759 2.73e-06 ***
v2           -5.0754     1.2290  -4.130  0.00167 **
v3          -12.9768     1.2056 -10.764 3.53e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.873 on 11 degrees of freedom
Multiple R-squared: 0.9403,     Adjusted R-squared: 0.9241
F-statistic: 57.78 on 3 and 11 DF,  p-value: 5.082e-07


> sum((y-mean(y))^2)
[1] 646.4
> sum((x-mean(x))^2)
[1] 360
> sum((x-mean(x))*y)
[1] 262
```

i. $Y = \beta_0 + \beta_1 x + \beta_2 v_2 + \beta_3 v_3 + \varepsilon.$

Then

Treatment 1 – $Y = \beta_0 \quad + \beta_1 x \quad + \varepsilon$

Treatment 2 – $Y = \beta_0 + \beta_2 \quad + \beta_1 x \quad + \varepsilon$

Treatment 3 – $Y = \beta_0 + \beta_3 \quad + \beta_1 x \quad + \varepsilon$

ii. $H_0: \beta_2 = \beta_3 = 0.$

iii. $Y = \beta_0 + \beta_1 x + \varepsilon.$

iv.

```
> fit0 = lm(y ~ x)
> anova(fit0,fit)
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ x + v2 + v3
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     13 455.72
2     11  38.57  2    417.15 59.483 1.264e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

0.000001264 = p-value < $\alpha = 0.05$. **Reject H_0**: $\beta_2 = \beta_3 = 0$ at $\alpha = 0.05$.

OR

i. $Y = \alpha_1 v_1 + \alpha_2 v_2 + \alpha_3 v_3 + \beta x + \varepsilon.$

ii. $H_0: \alpha_1 = \alpha_2 = \alpha_3.$

iii. $Y = \alpha + \beta x + \varepsilon.$

iv.

```
> fit2 = lm(y ~ v1 + v2 + v3 + x + 0)
> anova(fit0,fit2)
Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ v1 + v2 + v3 + x + 0
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     13 455.72
2     11  38.57  2    417.15 59.483 1.264e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

==

Full model: $Y = \beta_0 + \beta_1 x + \beta_2 v_2 + \beta_3 v_3 + \varepsilon.$

Need $\text{SSResid}_{\text{Full}}$.

$$s_e = \sqrt{\frac{\text{SSResid}}{n-p}} \quad \Rightarrow \quad 1.873 = \sqrt{\frac{\text{SSResid}}{15-4}} = \sqrt{\frac{\text{SSResid}}{11}}$$

$$\Rightarrow \quad \text{SSResid} \approx \mathbf{38.59} \qquad (\text{rounding})$$

OR

$$R^2 = 1 - \frac{\text{SSResid}}{\text{SSTotal}} \quad \Rightarrow \quad 0.9403 = 1 - \frac{\text{SSResid}}{646.4}$$

$$\Rightarrow \quad \text{SSResid} \approx \mathbf{38.59} \qquad (\text{rounding})$$

Null model: $Y = \beta_0 + \beta_1 x + \varepsilon.$ – simple linear regression.

Need $\text{SSResid}_{\text{Null}}$.

$$\hat{\beta}_1 = \frac{\text{SXY}}{\text{SXX}} = \frac{262}{360}$$

$$\text{SSRegr} = \hat{\beta}_1^2 \, \text{SXX} = \left(\frac{262}{360}\right)^2 \cdot 360 \approx 190.67778$$

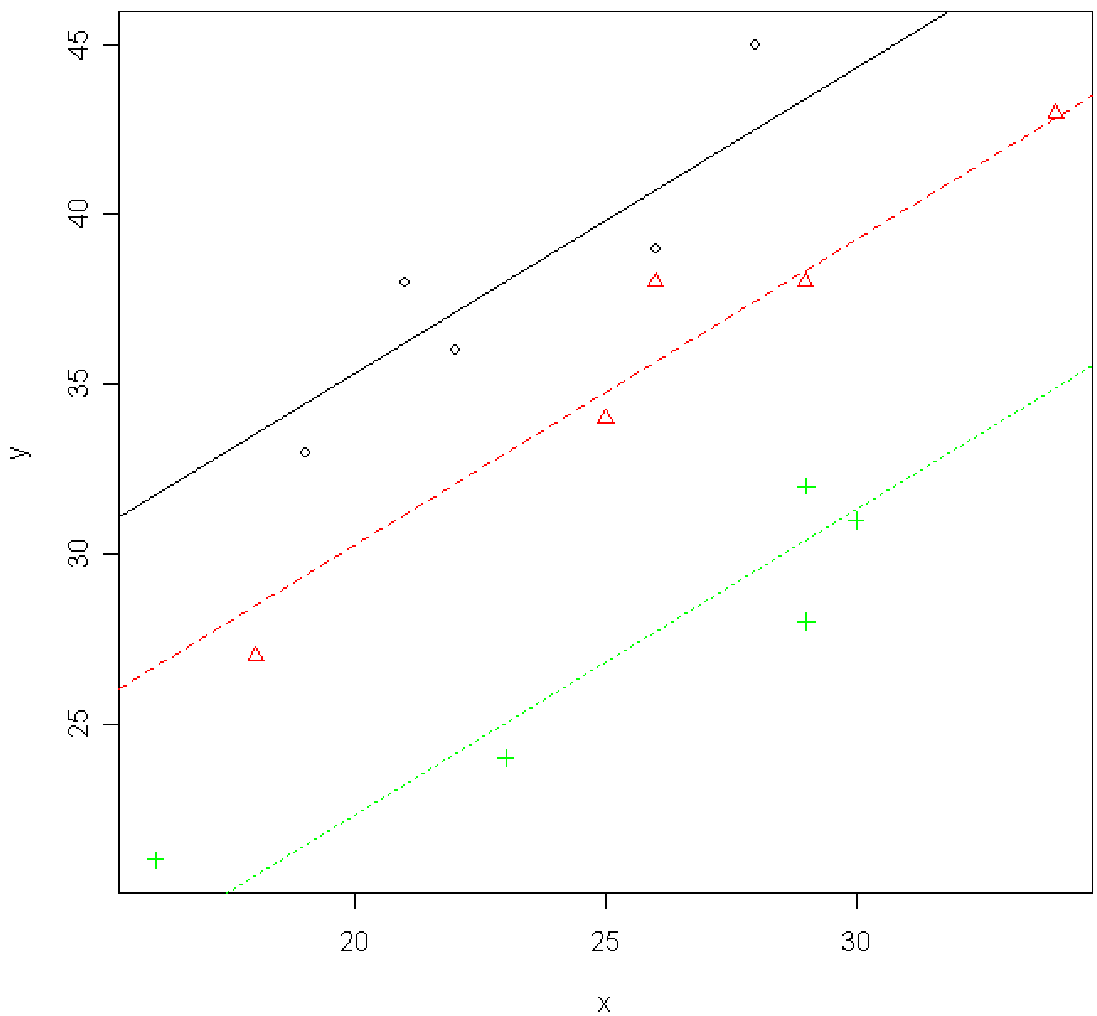
$$\text{SSResid} = \text{SYY} - \text{SSRegr} \approx 646.4 - 190.67778 = \mathbf{455.72222}.$$

Source	SS	DF	MS	F
H_0 (Diff.)	417.13	$4-2=2$	208.565	**59.45**
Full	38.59	$15-4=11$	3.508	
Null	455.72	$15-2=13$		

$F_{0.05}(2, 11) = 3.98$ **Reject H_0**: $\beta_2 = \beta_3 = 0$ at $\alpha = 0.05$.

$F_{0.01}(2, 11) = 7.21$ **Reject H_0**: $\beta_2 = \beta_3 = 0$ at $\alpha = 0.01$.

```
> plot(x,y,col=1+v2+2*v3,pch=1+v2+2*v3)
> abline(fit$coeff[1],fit$coeff[2],col=1,lty=1)
> abline(fit$coeff[1]+fit$coeff[3],fit$coeff[2],col=2,lty=2)
> abline(fit$coeff[1]+fit$coeff[4],fit$coeff[2],col=3,lty=3)
```



==
==
==

Consider the model that fits three regression lines (one for each treatment) that are not necessarily parallel. Test for parallel slopes:

i. Specify the full model.

ii. Specify the null hypothesis H_0 in the notations of your full model.

iii. Specify the model under the null hypothesis H_0.

iv. Conduct the *F* test at significance level $\alpha = 0.05$. (Show the calculations leading to your conclusion in the form of an ANOVA table.) State your decision/conclusion.

i. $Y = \beta_0 + \beta_1 x + \beta_2 v_2 + \beta_3 v_3 + \gamma_2 x v_2 + \gamma_3 x v_3 + \varepsilon,$

where v_2 is the indicator of Treatment 2, and v_3 is the indicator of Treatment 3.

Then

$$\begin{array}{llllll} \text{Treatment 1} - & Y = \beta_0 & & + \beta_1 x & & + \varepsilon \\ \text{Treatment 2} - & Y = \beta_0 & + \beta_2 & + (\beta_1 + \gamma_2) x & & + \varepsilon \\ \text{Treatment 3} - & Y = \beta_0 & + \beta_3 & + (\beta_1 + \gamma_3) x & & + \varepsilon \end{array}$$

ii. $H_0: \gamma_2 = \gamma_3 = 0.$

iii. $Y = \beta_0 + \beta_1 x + \beta_2 v_2 + \beta_3 v_3 + \varepsilon.$

iv.

```
> fit3 = lm(y ~ x + v2 + v3 + I(x*v2) + I(x*v3))
> anova(fit,fit3)
Analysis of Variance Table

Model 1: y ~ x + v2 + v3
Model 2: y ~ x + v2 + v3 + I(x * v2) + I(x * v3)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     11 38.571
2      9 31.521  2     7.050 1.0065 0.4032
```

$0.4032 = \text{p-value} > \alpha = 0.05$ **Do NOT Reject H_0**

OR

i. $Y = \alpha_1 v_1 + \alpha_2 v_2 + \alpha_3 v_3 + \beta_1 x v_1 + \beta_2 x v_2 + \beta_3 x v_3 + \varepsilon.$

ii. $H_0: \beta_1 = \beta_2 = \beta_3.$

iii. $Y = \alpha_1 v_1 + \alpha_2 v_2 + \alpha_3 v_3 + \beta x + \varepsilon.$

iv.

```
> fit4 = lm(y ~ v1 + v2 + v3 + I(x*v1) + I(x*v2) + I(x*v3) + 0)
> anova(fit2,fit4)
Analysis of Variance Table

Model 1: y ~ v1 + v2 + v3 + x + 0
Model 2: y ~ v1 + v2 + v3 + I(x * v1) + I(x * v2) + I(x * v3) + 0
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     11 38.571
2      9 31.521  2     7.050 1.0065 0.4032
```

0.4032 = p-value > α = 0.05 **Do NOT Reject H_0**