Math 3080 § 1. **Final Exam** Name:_____
Treibergs May 6, 2010

(1.) *In an experiment to see how hypertension is related to smoking habits, the following data was taken on 180 individuals. Test the hypothesis that the presence or absence of hypertension and the smoking habits are independent. Use a .05 level of significance. State the null hypothesis, the test statistic, the rejection region, your computation and conclusion. [Hint: the sum* = 14.46358.*]*

	Nonsmokers	Moderate Smokers	Heavy Smokers	Total
Hypertension	21	36	30	87
No Hypertension	48	26	19	93
Total	69	62	49	180

Let $p_{ij} = \mathbb{P}(A_i \cap B_j)$ denote the proportion of the population in the hypertension i and smoking j cell, $p_i = \mathbb{P}(A_i)$ the probability of hypertension i and $q_j = \mathbb{P}(B_j)$ the probability of smoker j. The null and alternative hypotheses are

$$\begin{aligned} \mathcal{H}_0 &: \ p_{ij} = p_i q_j, \qquad \text{for all } i = 1, \ldots, I \text{ and } j = 1, \ldots, J \\ \mathcal{H}_1 &: \ p_{ij} \neq p_i q_j, \qquad \text{for some } (i, j). \end{aligned}$$

If T is the total number of observations, R_i are the row totals and C_j are the column totals then the estimators for the proportions are $\hat{p}_i = R_i/T$ and $\hat{q}_j = C_j/T$. Then e_{ij} denoting the expected number in the (i, j) cell is given by $e_{ij} = T\hat{p}_i\hat{q}_j = R_i C_j/T$. The table of expected numbers in each cell is

e_{ij}	$j = 1$	$j = 2$	j=3
$i = 1$	$e_{11} = \frac{R_1 C_1}{T} = \frac{87 \cdot 62}{180} = 33.350$	$e_{12} = \frac{R_1 C_2}{T} = \frac{87 \cdot 49}{180} = 29.914$	$e_{13} = \frac{R_1 C_3}{T} = \frac{87 \cdot 69}{180} = 23.683$
$i = 2$	$e_{21} = \frac{R_2 C_1}{T} = \frac{93 \cdot 69}{180} = 35.650$	$e_{22} = \frac{R_2 C_2}{T} = \frac{93 \cdot 62}{180} = 32.033$	$e_{23} = \frac{R_2 C_3}{T} = \frac{93 \cdot 49}{180} = 25.317$

All expected cell counts exceed five, so we may use the χ^2 test. Under $\mathcal{H}_0$, it is asymptotically distributed as χ^2 with $(I-1)(J-1)$ degrees of freedom. The test statistic is

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(y_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2_{(I-1)(J-1)}.$$

where y_{ij} is the observed cell count. The null hypothesis is rejected if $\chi^2 > \chi^2_{(I-1)(J-1)}(\alpha) = \chi^2_2(.05) = 5.991$. The hint tells us that

$$\chi^2 = \frac{(21 - 33.350)^2}{33.350} + \cdots + \frac{(19 - 25.317)^2}{25.317} = 14.464.$$

Since this exceeds the critical value we reject $\mathcal{H}_0$: the data strongly indicates that the hypertension and smoking habits are not independent.

(2.) *Consider a one factor fixed effects ANOVA model with* $I = 3$ *and* $J = 2$

$$Y_{ij} = \mu_i + \epsilon_{ij}, \qquad \textit{for } i = 1, 2, 3 \textit{ and } j = 1, 2 \tag{1}$$

where μ_i *are constants and the* $\epsilon_{ij} \sim N(0, \sigma^2)$ *are IID normal random variables. Formulate the problems as a multiple regression* $y = X\beta + \epsilon$. *What are your* n, p, *the* $n \times p$ *design matrix*

$$X = \begin{pmatrix} x_{11} & \cdots & x_{p1} \\ \vdots & \ddots & \vdots \\ x_{1n} & \dots & x_{pn} \end{pmatrix},$$

and the $n \times 1$ *matrices* y *and* ϵ*? Using your* X *and* y, *find the estimator* $\hat{\beta}$. *Show how this gives the usual estimators for* $\hat{\mu}_i$.

There are $3 \cdot 2 = 6$ observations so $n = 6$. There are four parameteers to estimate: μ_1, μ_2, μ_3 and σ, thus $p = 3$. There are two ways to solve the problem. The easiest way is to set

$$y = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \qquad \epsilon = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}$$

where we are just coding columns of X for the categorical variable i at three levels. Thus the equation $y = X\beta + \epsilon$ is exactly (1). The estimator for the regression is given by $\hat{\beta} = (X'X)^{-1}X'y$ where X' is the transpose of X. Computing,

$$X'X = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}, \qquad (X'X)^{-1} = \frac{1}{2} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Thus

$$\hat{\beta} = (X'X)^{-1}X'y = \frac{1}{2}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix} = \begin{pmatrix} \frac{y_{11}+y_{12}}{2} \\ \frac{y_{21}+y_{22}}{2} \\ \frac{y_{31}+y_{32}}{2} \end{pmatrix} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \hat{\mu}_3 \end{pmatrix},$$

which are the usual estimators for the means $\hat{\mu}_i = \bar{y}_{i\cdot}$.

The other way follows how programs like **R**© convert ANOVA to regression. One sets

$$y = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}, \qquad \beta = \begin{pmatrix} \mu_1 \\ \mu_2 - \mu_1 \\ \mu_3 - \mu_1 \end{pmatrix}, \qquad \epsilon = \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{31} \\ \epsilon_{32} \end{pmatrix}$$

Thus the equation $y = X\beta + \epsilon$ is also equivalent to (1).

The estimator for the regression is given by $\hat{\beta} = (X'X)^{-1}X'y$ where X' is the transpose of X. Computing,

$$X'X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 6 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{pmatrix}, \qquad (X'X)^{-1} = \frac{1}{2}\begin{pmatrix} 1 & -1 & -1 \\ -1 & 2 & 1 \\ -1 & 1 & 2 \end{pmatrix}.$$

Thus

$$\hat{\beta} = (X'X)^{-1}X'y = \frac{1}{2}\begin{pmatrix} 1 & -1 & -1 \\ -1 & 2 & 1 \\ -1 & 1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}\begin{pmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix}$$

$$= \frac{1}{2}\begin{pmatrix} 1 & -1 & -1 \\ -1 & 2 & 1 \\ -1 & 1 & 2 \end{pmatrix}\begin{pmatrix} y_{11}+y_{12}+y_{21}+y_{22}+y_{31}+y_{32} \\ y_{21}+y_{22} \\ y_{31}+y_{32} \end{pmatrix}$$

$$= \frac{1}{2}\begin{pmatrix} y_{11}+y_{12} \\ -y_{11}-y_{12}+y_{21}+y_{22} \\ -y_{11}-y_{12}+y_{31}+y_{32} \end{pmatrix} = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2-\hat{\mu}_1 \\ \hat{\mu}_3-\hat{\mu}_3 \end{pmatrix},$$

which implies that $\hat{\mu}_i = \bar{y}_{i\cdot}$ which are the usual estimators for the means .

(3.) *A simple regression study in R. A. Johnson's, Probability and Statistics for Engineers reports data on percent carbon content (X) and permeability index (Y) for 22 sinter mixtures. Data pairs and partial* **R©** *output is given. If another reading were made at $X^* = 4$, what do you predict that the corresponding expected permeability index $\mathbb{E}(Y^*)$ will be? Give an .05 prediction interval when $X^* = 4$. [Hint: $\bar{x} = 4.655$, $\bar{y} = 19.182$, $S_{xx} = 6.675$, $S_{xy} = -43.318$, $S_{yy} = 801.273$.] State the hypotheses for the model. Comment on how well they are satisfied.*

```
X  4.4 5.5 4.2 3.0 4.5 4.9 4.6 5.0 4.7 5.1 4.4 4.1 4.9 4.7 5.0 4.6 3.6 4.9 5.1 4.8 5.2 5.2
Y   12  14  18  35  23  29  16  12  18  21  27  13  19  22  20  16  27  21  13  18  17  11

Call:  f1 <- lm(formula = Y ~ X)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   48.693      9.365   5.199 4.35e-05
X             -6.340      1.998  -3.173  0.00478

Residual standard error: 5.162 on 20 degrees of freedom
Multiple R-squared: 0.3349,Adjusted R-squared: 0.3016

Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value   Pr(>F)
X          1 268.31 268.307  10.069 0.004781
Residuals 20 532.97  26.648
```

```
Shapiro-Wilk normality test

data:  resid(f1)
W = 0.978, p-value = 0.8828
```

If we wish to predict the expected y when $x^* = 4$, we use the predicting line from the output

$$\mathbb{E}(y^*) = \hat{\beta}_0 + \hat{\beta}_1 x^* = 48.693 + (-6.340)(4.000) = 23.333.$$

The $\alpha = .05$ level prediction interval, using the hints and output

$$\begin{aligned}\mathbb{E}(y^*) \pm t_{n-2}(\alpha) s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} &= 23.333 \pm (2.086)(5.162)\sqrt{1 + \frac{1}{22} + \frac{(4 - 4.655)^2}{6.675}} \\ &= 23.333 \pm 11.346 = (11.987, 34.679).\end{aligned}$$

The hypotheses for the simple regression model is that the x_i are assumed to be known constants and Y_i are random variables such that for all observations $i = 1, \ldots, n$,

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$ are IID normal random variables. Without the diagnostic plots, we can only address whether the randomness in the data is in fact normal. The residuals $e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$, which are linear combinations of the y_i should also be normally distributed. Indeed, the Shapiro-Wilk test for the normality of the residuals gives a P-value of .8828, so there is little evidence that normality is violated.

(4.) *Consider a one factor fixed effects ANOVA model*

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad \textit{for } i = 1, \ldots, I \textit{ and } j = 1, \ldots, J$$

where μ and α_i are constants such that $\sum_i \alpha_i = 0$ and $\epsilon_{ij} \sim N(0, \sigma^2)$ are IID normal random variables. Find the expectation $\mathbb{E}(Z)$ of the random variable

$$Z = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2$$

The random variable is of course $Z = SSE$, the sum squared error. First, we square the summands and replace SS by the computation formula

$$SSE = \sum_{i=1}^{I} \sum_{j=1}^{J} Y_{ij}^2 - J \sum_{i=1}^{I} \bar{Y}_{i\cdot}^2.$$

The expected values of the sum of squares uses the formula for variance of a random variable X, namely $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)$. Thus, using independence,

$$\begin{gathered}\mathbb{E}(Y_{ij}) = \mathbb{E}\left(\mu + \alpha_i + \epsilon_{ij}\right) = \mu + \alpha_i, \\ \mathbb{E}(Y_{ij}^2) = \mathbb{V}(Y_{ij}) + \mathbb{E}^2(Y_{ij}) = \mathbb{V}\left(\mu + \alpha_i + \epsilon_{ij}\right) + (\mu + \alpha_i)^2 = \sigma^2 + (\mu + \alpha_i)^2, \\ \bar{Y}_{i\cdot} = \frac{1}{J}\sum_{j=1}^{J} Y_{ij} = \frac{1}{J}\sum_{j=1}^{J} \left(\mu + \alpha_i + \epsilon_{ij}\right) = \mu + \alpha_i + \frac{1}{J}\sum_{j=1}^{J} \epsilon_{ij} \\ \mathbb{E}(\bar{Y}_{i\cdot}) = \mathbb{E}\left(\mu + \alpha_i + \frac{1}{J}\sum_{j=1}^{J} \epsilon_{ij}\right) = \mu + \alpha_i, \\ \mathbb{E}(\bar{Y}_{i\cdot}^2) = \mathbb{V}\left(\mu + \alpha_i + \frac{1}{J}\sum_{j=1}^{J} \epsilon_{ij}\right) + \mathbb{E}^2(\bar{Y}_{i\cdot}) = \frac{1}{J}\sigma^2 + (\mu + \alpha_i)^2\end{gathered}$$

Putting these together we have the formula for the expectations of the sums of squares

$$\begin{aligned}
\mathbb{E}(SSE) &= \mathbb{E}\left(\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}^2 - J\sum_{i=1}^{I}\bar{Y}_{i\cdot}^2\right) \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\mathbb{E}\left(\bar{Y}_{ij}^2\right) - J\sum_{i=1}^{I}\mathbb{E}\left(\bar{Y}_{i\cdot}^2\right) \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\left\{\sigma^2 + (\mu+\alpha_i)^2\right\} - J\sum_{i=1}^{I}\left\{\frac{1}{J}\sigma^2 + (\mu+\alpha_i)^2\right\} \\
&= I(J-1)\sigma^2.
\end{aligned}$$

(5.) The paper "...Protocols for Mobile Ad Hoc Networks," *Proceedings 2002 International Conference on Wireless Networks*, tried to predict network performance measured by y data overhead (in kB) in terms of x_1 speed of computers (m/s), x_2 pause time at each link (s) and x_3 the link change rate (100/s). Consider fitting the full quadratic model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1^2 + \beta_8 x_2^2 + \beta_9 x_3^2 + \epsilon$. Here is the data and **R** output of the analysis of variance. Can you conclude that $\beta_2 < -2$? Perform the appropriate hypothesis test. If you were looking to take a step to improve this model, what variable(s) would you consider dropping from the model? Which would you keep? Why?

Speed	Pause	LCR	Overhead	Speed	Pause	LCR	Overhead	Speed	Pause	LCR	Overhead
5	10	9.43	428.90	10	50	8.31	498.77	30	30	16.70	506.23
5	20	8.32	443.68	20	10	26.31	452.24	30	40	13.26	516.27
5	30	7.37	452.38	20	20	19.01	475.97	30	50	11.11	508.18
5	40	6.74	461.24	20	30	14.73	499.67	40	10	37.82	444.41
5	50	6.06	475.07	20	40	12.12	501.48	40	20	24.14	490.58
10	10	16.46	446.06	20	50	10.28	519.20	40	30	17.70	511.35
10	20	13.28	465.89	30	10	33.01	445.45	40	40	14.06	523.12
10	30	11.16	477.07	30	20	22.13	489.02	40	50	11.69	523.36
10	40	9.51	488.73								

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 435.99048   25.82529  16.882 3.62e-11
X1            0.56556    2.35353   0.240  0.81335
X2           -2.15504    1.29222  -1.668  0.11611
X3           -2.24927    3.26020  -0.690  0.50078
X1X2         -0.04820    0.03145  -1.533  0.14616
X1X3         -0.14612    0.08428  -1.734  0.10347
X2X3          0.36358    0.09438   3.853  0.00157
X1X1          0.05117    0.02558   2.001  0.06386
X2X2          0.02362    0.01292   1.828  0.08754
X3X3          0.07581    0.09187   0.825  0.42222

Residual standard error: 4.205 on 15 degrees of freedom
Multiple R-squared: 0.9868,Adjusted R-squared: 0.9789
F-statistic: 124.8 on 9 and 15 DF,  p-value: 1.863e-12

Analysis of Variance Table.   Response: Y
          Df  Sum Sq Mean Sq  F value    Pr(>F)
X1         1  5571.2  5571.2 315.0457 1.766e-11
X2         1 10973.9 10973.9 620.5618 1.283e-13
X3         1   558.8   558.8  31.5973 4.870e-05
```

```
X1X2        1     0.1     0.1   0.0066    0.9362
X1X3        1  2073.4  2073.4 117.2461 1.737e-08
X2X3        1   585.4   585.4  33.1010 3.814e-05
X1X1        1    32.0    32.0   1.8106    0.1984
X2X2        1    52.3    52.3   2.9577    0.1060
X3X3        1    12.0    12.0   0.6809    0.4222
Residuals 15   265.3    17.7
```

We use a one sided t-test to test

$$\begin{aligned} \mathcal{H}_0 &: \beta_2 \geq 0 \qquad \text{vs.} \\ \mathcal{H}_1 &: \beta_2 < 0. \end{aligned}$$

There are $n = 25$ observations and the number of β's to fit is $p = 10$. The test statistic is distributed as t with $n - p$ degrees of freedom.

$$T = \frac{\beta_2 - 0}{s(\beta_2)} \sim t_{n-p}.$$

The null hypothesis is rejected if $T < -t_{n-p}(\alpha) = -t_{25-10}(\alpha)$. However, this statistic is already computed in the output to be $T = -1.668$ but the P-values there are two sided so don't apply to this test. We reject the null hypothesis: there is mild evidence that $\beta_2 < 0$ at the $\alpha = .10$ level since $t_{15}(.10) = 1.341$ but it is not significant at the $\alpha = .05$ level since $t_{15}(.05) = 1.753$.

In deciding what variables to remove in the next regression run, we look at the interactions of higher order term that are plausibly zero, namely the x_3^2 and x_1x_2 terms, whose coefficients have the highest p-values for being zero. I would keep x_1, x_2, x_3, x_2x_3, and x_1^2 because x_2x_3 and x_1^2 have much lower p-values, and one keeps first order terms of any significant interacting variables. The x_1x_3 has large p value too, but it is significant in the ANOVA table so it can be kept in the model another step.

(6.) *Consider a two factor fixed effects ANOVA model*

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \qquad \textit{for } i = 1, \ldots, I,\ j = 1, \ldots, J \textit{ and } k = 1, \ldots, K,$$

where μ_{ij} is a constant and $\epsilon_{ijk} \sim N(0, \sigma^2)$ is an IID normal random variable. $\hat{\mu}_{ij}$ is chosen to be least squares estimator, which means that it minimizes a certain sum of squares. Give a formula for this sum of squares. Minimize your sum of squares to deduce the formula for $\hat{\mu}_{ij}$.

The least squares estimators are chosen to minimize the least square errors. For a given choice of esimators $\hat{\mu}_{ij}$, the corresponding SSE is

$$\mathcal{Q}(\hat{\mu}) = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \left(\hat{\mu}_{ij} - y_{ijk}\right)^2,$$

where $\hat{\mu} = (\ldots, \hat{\mu}_{ij}, \ldots)$ is the $I \times J$ matrix of $\hat{\mu}_{ij}$'s. There are two ways to find $\hat{\mu}$ that minimizes $\mathcal{Q}$: by setting the derivatives equal to zero or by "completing the square." Using the first way, let us fix one of the (i_0, j_0) with $1 \leq i_0 \leq I$ and $1 \leq j_0 \leq J$ and take the derivative with respect to $\hat{\mu}_{i_0j_0}$. Splitting the sum into terms that involve $\hat{\mu}_{i_0j_0}$ and those that don't and differentiating,

$$\begin{aligned} \frac{\partial\, \mathcal{Q}}{\partial \hat{\mu}_{i_0j_0}} &= \frac{\partial}{\partial \hat{\mu}_{i_0j_0}} \left(\sum_{k=1}^{K} \left(\hat{\mu}_{i_0j_0} - y_{i_0j_0k}\right)^2 + \sum_{(i,j) \neq (i_0, j_0)} \sum_{k=1}^{K} \left(\hat{\mu}_{ij} - y_{ijk}\right)^2 \right) \\ &= 2 \sum_{k=1}^{K} \left(\hat{\mu}_{i_0j_0} - y_{i_0j_0k}\right) + 0 = 2K\hat{\mu}_{i_0j_0} - 2 \sum_{k=1}^{K} y_{i_0j_0k}. \end{aligned}$$

These expressions are all zero if and only if for each (i_0, j_0),

$$\hat{\mu}_{i_0 j_0} = \frac{1}{K}\sum_{k=1}^{K} y_{i_0 j_0 k} = \bar{y}_{i_0 j_0 \bullet} \tag{2}$$

The second way is to split the sum of squares

$$\begin{aligned}
\mathcal{Q} &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(\hat{\mu}_{ij} - \bar{y}_{ij\cdot} + \bar{y}_{ij\cdot} - y_{ijk}\right)^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(\hat{\mu}_{ij} - \bar{y}_{ij\cdot}\right)^2 + 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(\hat{\mu}_{ij} - \bar{y}_{ij\cdot}\right)\left(\bar{y}_{ij\cdot} - y_{ijk}\right) + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(\bar{y}_{ij\cdot} - y_{ijk}\right)^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(\hat{\mu}_{ij} - \bar{y}_{ij\cdot}\right)^2 + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(\bar{y}_{ij\cdot} - y_{ijk}\right)^2
\end{aligned}$$

which is minimized when the first sum is zero by choosing (2) for each (i_0, j_0). The cross terms vanish because

$$\begin{aligned}
\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(\hat{\mu}_{ij} - \bar{y}_{ij\cdot}\right)\left(\bar{y}_{ij\cdot} - y_{ijk}\right) &= \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\hat{\mu}_{ij} - \bar{y}_{ij\cdot}\right)\sum_{k=1}^{K}\left(\bar{y}_{ij\cdot} - y_{ijk}\right) \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\hat{\mu}_{ij} - \bar{y}_{ij\cdot}\right)\left(K\bar{y}_{ij\cdot} - K\bar{y}_{ij\cdot}\right) = 0.
\end{aligned}$$

(7.) *In the study "Vitamin C Retention in Reconstituted Frozen Orange Juice," (VPI Department of Human Nutrition and Foods, 1972), three brands (R =Richfood, S =Sealed-Sweet, M =Minute Maid) were measured at three different time periods (0,3,7 days) between when OJ concentrate was blended and when it was tested. Response is mg/l ascorbic acid. Here is the data and partial SAS output. State the assumptions of the model. Test for interactions of the main effects. State the null hypotheses, test statistic, rejection region and your conclusion. Compute Tukey's HSD_{time} using $\alpha = .05$. Using Tukey's HSD, determine which pairs of time means are significantly different.*

	Time = 0 days				3 days				7 days			
Brand												
R	52.6	54.2	49.8	46.5	49.4	49.2	42.8	53.2	42.7	48.8	40.4	47.6
S	56.0	48.0	49.6	48.4	48.8	44.0	44.0	42.4	49.2	44.0	42.0	43.2
M	52.5	52.0	51.8	53.6	48.0	47.0	48.2	49.6	48.5	43.3	45.2	47.6

The GLM Procedure

Dependent Variable: Acid Ascorbic-Acid

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	277.2850000	34.6606250	3.67	0.0051
Error	27	254.7025000	9.4334259		
Corrected Total	35	531.9875000			

R-Square	Coeff Var	Root MSE	acid Mean
0.521225	6.413200	3.071388	47.89167

Source	DF	SS	Mean Square	F Value	Pr > F
Brand	2	32.7516667	16.3758333	1.74	0.1953
Time	2	227.2116667	113.6058333	12.04	0.0002
Brand*Time	4	17.3216667	4.3304167	0.46	0.7650

Level of Brand	N	Acid Mean	Acid Std Dev
M	12	48.9416667	3.09470467
R	12	48.1000000	4.35994162
S	12	46.6333333	4.09863244

Level of Time	N	Acid Mean	Acid Std Dev
0	12	51.2500000	2.81408794
3	12	47.2166667	3.27131704
7	12	45.2083333	3.01434701

In this study, both factors have $I = J = 3$ levels and there are $K = 4$ replications. We are assuming a two factor fixed effects ANOVA with interactions. Thus we assume that the sample consists of random variables of the form

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \qquad \text{for all } i = 1, \ldots, I,\ j = 1, \ldots, J \text{ and } k = 1, \ldots, K,$$

where μ, α_i, β_j and $(\alpha\beta)_{ij}$ are constants such that $\sum_i \alpha_i = \sum_j \beta_j = 0$, $\sum_i (\alpha\beta)_{ij} = 0$ for all j, $\sum_j (\alpha\beta)_{ij} = 0$ for all i and $\epsilon_{ijk} \sim N(0, \sigma^2)$ are independent, identically distributed normal random variables.

We test for the presence of interactions. The null and alternative hypotheses are

$$\begin{aligned} & \mathcal{H}_0 : (\alpha\beta)_{ij} = 0 \qquad \text{for all } i = 1, \ldots, I \text{ and } j = 1, \ldots, J. \\ \text{vs.} \quad & \mathcal{H}_1 : (\alpha\beta)_{ij} \neq 0 \qquad \text{for some } (i, j). \end{aligned}$$

The test statistic is $F_{AB} = MSAB/MSE \sim f_{(I-1)(J-1),IJ(K-1)}$ which is distributed as an f variable with $\nu_1 = (I-1)(J-1)$ and $\nu_2 = IJ(K-1)$ degrees of freedom. The null hypothesis is rejected at the $\alpha = .05$ level if $F_{AB} > f_{(I-1)(J-1),IJ(K-1)}(\alpha) = f_{4,27}(.05) = 2.73$. In this case, $F_{AB} = 0.46$ with p-value of .7650 so that $\mathcal{H}_0$ is not rejected: the interaction terms are plausibly zero.

On the other hand, the p-value for the time factor is .0002 which is highly significant: there is statistical evidence that the β_j are not zero. We compute Tukey's honest significant difference. $n = J = 3$, the number of means compared and $\nu = IJ(K-1) = 27$ is the degrees of freedom in the MSE term. As $\bar{Y}_{\cdot j\cdot}$ is an average over IK terms,

$$HSD_{\text{time}} = q(\alpha; n, \nu)\sqrt{\frac{MSE}{IK}} = q(.05, 3, 27)\sqrt{\frac{9.4334259}{12}} = 3.51\sqrt{\frac{9.4334259}{12}} = 3.11.$$

Note that the Studentized Range is not given for $(n, \nu) = (3, 27)$ so we interpolated the table: since $27 = .5(24 + 30)$ we use the straight line approximation $q(.05, 3, 27) \approx .5(q(.05, 3, 24) + q(.05, 3, 30)) = .5(3.53 + 3.49) = 3.51$. Or use the conservative value 3.53. Computing the differences we find

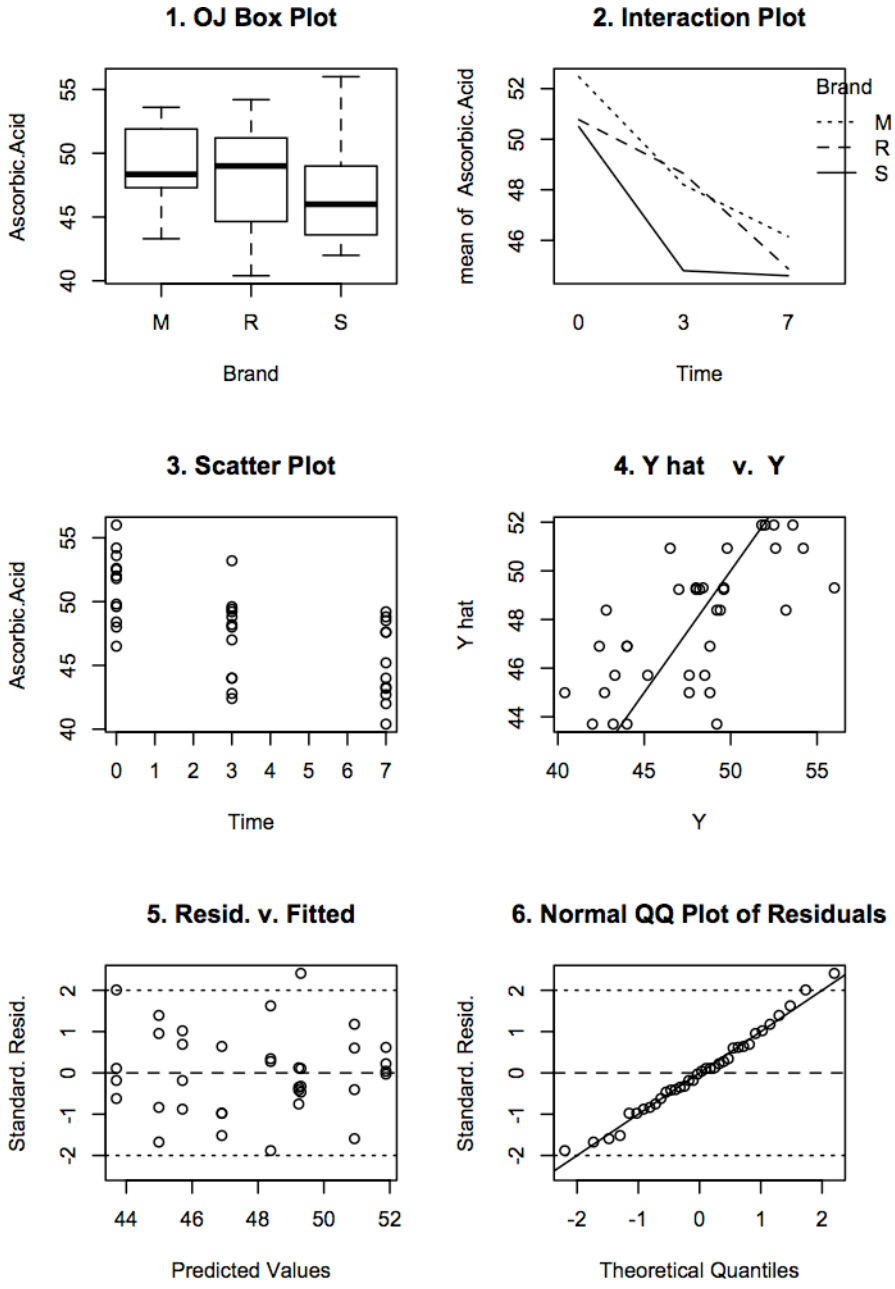
$$\begin{aligned} \bar{Y}_{1\cdot} - \bar{Y}_{2\cdot} &= 51.250 - 47.217 = 4.033, \\ \bar{Y}_{1\cdot} - \bar{Y}_{3\cdot} &= 51.250 - 45.208 = 6.042, \\ \bar{Y}_{2\cdot} - \bar{Y}_{3\cdot} &= 47.217 - 45.208 = 2.009 \end{aligned}$$

The first two exceed HSD and are significant, the third does not. The Tukey Bar pattern is thus

$$\underline{51.250 \qquad 47.217} \qquad 45.208.$$

The acid at time 0 days is significantly greater than the acid at 3 days or 7 days. However the acid levels at 3 days and 7 days were not significantly different.

(8.) *The same data from the study "Vitamin C Retention in Reconstituted Frozen Orange Juice," as in Problem 7 was used to produce six diagnostic plots in* **R**©. *For each of the six plots shown, briefly explain what information about the data, the analysis or the appropriateness of the model can be concluded from that plot.*



Plot 1. The boxplots indicate the median and spread of OJ vitamin C (Acid) for different brands. Here, the inter-quartile differences (box heights) are nearly equal indicating that the spread of Acid is roughly uniform with respect to Brand, which upholds the hypothesis that spread be independent of the variables. The median Acid for Brand S is a bit smaller than the others.

Plot 2. The interaction plots show how the mean Acid changes in time for the different Brands. Here, as all three are decreasing, the lines are roughly but not perfectly parallel, indicating that the interaction term is small but nonzero.

Plot 3. The scatter plots indicate the spread of Acid for different times. Here the spreads are nearly equal indicating that the spread of Acid is roughly uniform with respect to Time, which upholds the hypothesis that spread be independent of the variables. The Acid levels drop as time increases.

Plot 4. Here, the $\hat{Y}$ v. Y plot shows uniform horizontal spread independent of $\hat{Y}$, indicating that the variance does not depend on the predicted value, which is assumed by the model. On the other hand, the points are not near the $\hat{Y} = Y$ line, indicating that the model does not predict the outcome well. This plot is a way check model effectiveness in making predictions.

Plot 5. The plot of standardized residuals v. fitted values is an important one to see if model hypotheses are satisfied. The residuals and fitted values are independent so that the scatter should be uniform ("tubular") over the range of predicted values. In this case, the vertical spread is uniform as expected from our model assumptions. The spread has been standardized (residuals have been divided by the standard error) which gives $N(0,1)$ variable if the assumptions are met. That means that about 95% of the residuals should be within ± 2 standard deviations of zero. With 36 data points, seeing the one point outside the ± 2 dotted lines does not upset us.

Plot 6. The normal QQ-plot of standardized residuals tells about the distribution of residuals. It would look the same without standardizing. Under the model hypotheses, the errors, thus also the residuals should distribute as a normal variable. The observed quantiles are plotted against the theoretically normal quantiles. Failure of normality will look like bowing (skewed data) or "N/S" shaped (kurtotic data). Here the points align nicely with the 45° line, indicating normality hypotheses not being violated.

(9.) *The study "Split Plot Designs...for Mixture Experiments with Process Variables," (Technometrics, 2002) considered a* 2^3 *with four replicates design to study how the factors A proportion of plasticizer, B rate of extrusion and C drying temperature affect the thickness (in mils) in the manufacture of vinyl seat covers. Here is the printout of the data and the contrasts. Show by doing the computation that the contrast* $L_{ac} = 29$ *is correct. State the hypothesis of your model. Construct the ANOVA table.* $\left[\textit{Hint: } \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{\ell=1}^{4} y_{ijk\ell}^2 = 1655.\right]$

	Thickness				Total	Contrast
1	7	5	6	7	25	221
a	6	5	5	5	21	27
b	8	8	4	6	26	13
ab	9	5	6	9	29	11
c	7	6	5	5	23	19
ac	7	7	11	10	35	29
bc	6	4	5	8	23	-5
abc	8	11	11	9	39	-3

There are $n = 2^3 \cdot 4 = 32$ observations. Their sum is $L_1 = (1) + (a) + \cdots + (abc) = 221$. E.g., $(1) = 7 + 5 + 6 + 7$ is the total of replications under the experimental condition "1" (all factors at low level). The contrast $L_{\rm ac}$ is the inner product with the signs of the experimental condition $e_{\rm ac} = (1, -1, +1, -1, -1, 1, -1, 1)$, or

$$L_{\rm ac} = (1) - (a) + (b) - (ab) - (c) + (ac) - (bc) + (abc) = 25 - 21 + 26 - 29 - 23 + 35 - 23 + 39 = 29.$$

By the computation formula,

$$SST = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{\ell=1}^{4} y_{ijk\ell}^2 - \frac{1}{n}\left(\sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{\ell=1}^{4} y_{ijk\ell}\right)^2 = 1655 - \frac{221^2}{32} = 128.719.$$

The problem is shorter or longer, depending on what you choose the model to be. The longer computation results from the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijk\ell},$$

for all $i, j, k \in \{1, 2\}$ and $\ell = 1, \ldots, 4$ where μ, α_i, β_j, γ_k, $(\alpha\beta)_{ij}$, $(\alpha\gamma)_{ik}$, $(\beta\gamma)_{jk}$, $(\alpha\beta\gamma)_{ijk}$ are constants such that $\sum_i \alpha_i = \sum_j \beta_j = \sum_k \gamma_k = 0$, $\sum_k (\alpha\beta\gamma)_{ijk} = \sum_k (\beta\gamma)_{jk} = \sum_k (\alpha\gamma)_{ik} = 0$ for all i, j, $\sum_j (\alpha\beta\gamma)_{ijk} = \sum_j (\beta\gamma)_{jk} = \sum_j (\alpha\beta)_{ij} = 0$ for all i, k, $\sum_i (\alpha\beta\gamma)_{ijk} = \sum_i (\alpha\gamma)_{ik} = \sum_i (\alpha\beta)_{ij} = 0$ for all j, k and $\epsilon_{ijk\ell} \sim N(0, \sigma^2)$ are independent, identically distributed normal random variables. In the 2^3 design case, the sum squares are given by the formula

$$SS_{\text{factor}} = \frac{L_{\text{factor}}^2}{2^3 n}$$

E.g., $SSA = L_{\text{a}}^2/32 = 27^2/32 = 22.781$. Each of the seven SS have one degree of freedom. Then subtracting gives the residual sum of squares

$$\begin{aligned} SSE &= SST - SSA - SSB - SSAB - SSC - SSAC - SSBC - SSABC \\ &= 128.719 - 22.781 - \ldots - .281 = 58.250 \end{aligned}$$

with $n-1-7 = 24$ degrees of freedom. $MSE = SSE/DFE$. Then the $F_{\text{factor}} = MS_{\text{factor}}/MSE$. Here is the ANOVA table.

SOURCE	DF	SS	MS	F
a	1	22.781	22.781	9.386
b	1	5.281	5.281	2.176
ab	1	3.781	3.781	1.558
c	1	11.281	11.281	4.648
ac	1	26.281	26.281	10.828
bc	1	.781	.781	.322
abc	1	.281	.281	.116
Error	24	58.250	2.427	
Total	31	128.719		

The shorter answer is that you choose the additive model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk\ell},$$

for all $i, j, k \in \{1, 2\}$ and $\ell = 1, \ldots, 4$ where μ, α_i, β_j, γ_k are constants such that $\sum_i \alpha_i = \sum_j \beta_j = \sum_k \gamma_k = 0$ and $\epsilon_{ijk\ell} \sim N(0, \sigma^2)$ are independent, identically distributed normal random variables. Then subtracting gives the residual sum of squares

$$SSE = SST - SSA - SSB - SSC = 128.219 - 22.781 - 5.281 - 11.281 = 89.375$$

with $n - 1 - 3 = 28$ degrees of freedom. Here is the shorter ANOVA table.

SOURCE	DF	SS	MS	F
a	1	22.781	22.781	7.137
b	1	5.281	5.281	1.659
c	1	11.281	11.281	3.534
Error	28	89.375	3.192	
Total	31	128.719		

Since the critical $f_{1,24}(.05) = 4.26$, the AC term is significant in the long model. The A and C are significant too, but don't have simple interpretation due to the presence of the AC interaction. The critical $f_{1,28}(.05) = 4.20$ so that this time A is significant but C is not for the short model. But this is under the assumption that the interactions vanish.