Math 3080 § 1. **Second Midterm Exam** Name: Solutions
Treibergs March 31, 2010

1. Toughness and fibrousness of asparagus are major determinants of quality. The study "Post-Harvest Glyphosphate Application Reduces Toughening, Fiber Content, and Lignification of Stored Asparagus Spears" (*J. Amer. Soc. Horticultural Science*, 1988) reported the following data on shear force x (kg) and y percent fiber dry weight. There were $n = 18$ observations. Carry out a test at significance level .01 to decide whether there is a POSITIVE (y increases with x) linear association between the two variables.

$$\bar{x} = 108.3 \quad \bar{y} = 2.819$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = 40720 \quad S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 340. \quad S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 3.03$$

   (a) State your assumptions. State the null and alternative hypotheses.

   (b) State the test statistic, the rejection region.

   (c) Compute and describe your conclusions.

(a.) We assume that the points are a random sample $\{(x_i, y_i)\}$ from a bivariate normal distribution in $\mathbf{R}^2$. The null and alternative hypotheses to test if the data indicates whether there is a positive relationship are tests on the correlation coefficient ρ:

$$\begin{aligned} \mathcal{H}_0 &: \quad \rho = 0; \\ \mathcal{H}_1 &: \quad \rho > 0. \end{aligned}$$

(b.) The test statistic is Pearson's T, given by

$$T = R\sqrt{\frac{n-2}{1-R^2}} \qquad \text{where} \qquad R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}};$$

Under the null hypothesis $\rho = 0$, Pearson's T is distributed according to t_{n-2}, so we reject the null hypothesis if $T > t_{n-2}(\alpha)$ (one-sided test!) In this case $t_{16}(.01) = 2.583$.
(c.) We compute the sample correlation coefficient

$$R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{340.}{\sqrt{(40720)(3.03)}} = .9679518$$

so

$$T = R\sqrt{\frac{n-2}{1-R^2}} = \frac{.9679518\sqrt{16}}{\sqrt{1-(.9679518)^2}} = 15.42.$$

Thus we reject the null hypothesis: the data strongly indicates that there is a positive linear relationship between x and y.

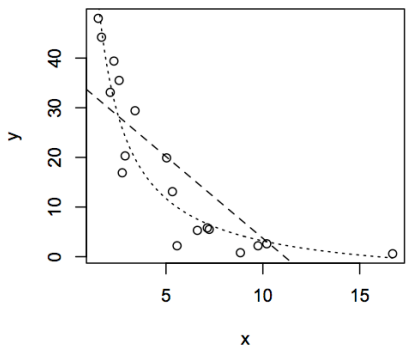
2. The paper "Oxidation State and Activities of Chromium Oxides...," (*Metallurgical and Materials Transactions B*, 2002) studied activity of chromium oxides y (activity coefficient) in terms of amount x (mole pct.) Consider fitting two models to explain the data: the first model is a simple linear regression $y = \beta_0 + \beta_1 x$, and the second is $y = \beta_0 + \beta_1/x$. **R** was used to generate the tables, (x, y) scatterplot, studentized residuals vs $\hat{y}$, $\hat{y}$ vs y and a normal PP-plot of studentized residuals. The top four panels show the first model. The bottom four show the second model.

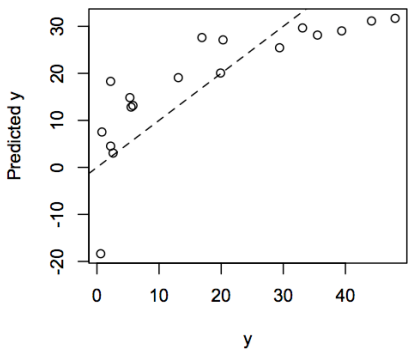
```
x  10.20   5.03   8.84   6.62   2.89   2.31   7.13   3.40   5.57
y   2.6   19.9    0.8    5.3   20.3   39.4    5.8   29.4    2.2

x   7.23   2.12   1.67   5.33  16.70   9.75   2.74   2.58   1.50
y   5.5   33.1   44.2   13.1    0.6    2.2   16.9   35.5   48.0
```

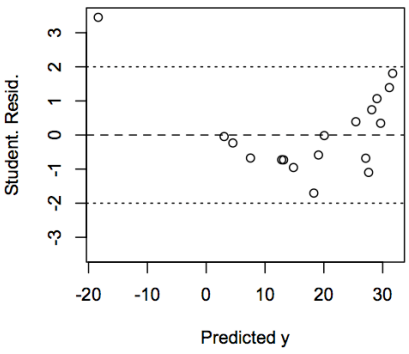
(a.) What does the second model predict to be the mean y when the mole pct. is 6.0?

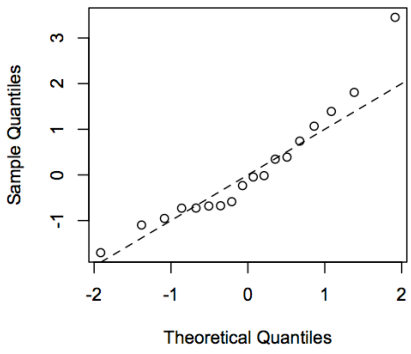
(b.) Discuss the two models with regard to quality of fit and whether model assumptions are satisfied. Compare at least five features of the models in the tables and plots. Which is the better model?

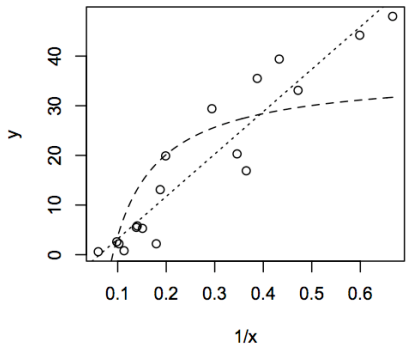
```
Call:
lm(formula = y ~ x)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.6320     4.3206   8.478 2.59e-07
x            -3.2927     0.6337  -5.196 8.83e-05

Residual standard error: 10.28 on 16 degrees of freedom
Multiple R-squared: 0.6279,Adjusted R-squared: 0.6046
F-statistic:    27 on 1 and 16 DF,  p-value: 8.831e-05

Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
x          1 2853.3 2853.27  27.000 8.831e-05
Residuals 16 1690.9  105.68

Call:
lm(formula = y ~ xinv)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -5.366      2.398  -2.237   0.0398
xinv          85.439      7.367  11.598 3.36e-09

Residual standard error: 5.495 on 16 degrees of freedom
Multiple R-squared: 0.8937,Adjusted R-squared: 0.8871
F-statistic: 134.5 on 1 and 16 DF,  p-value: 3.365e-09

Analysis of Variance Table
Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)
xinv       1 4061.1  4061.1  134.52 3.365e-09
Residuals 16  483.0    30.2
```

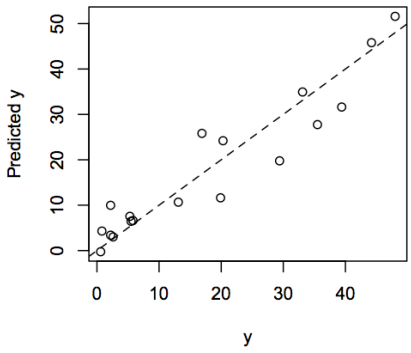

y
x

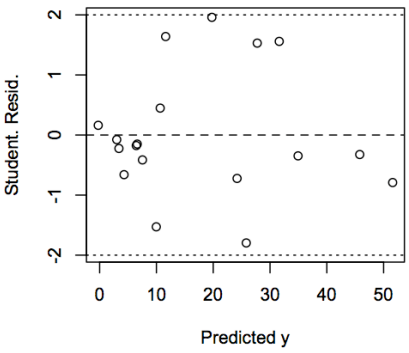

Predicted y
y

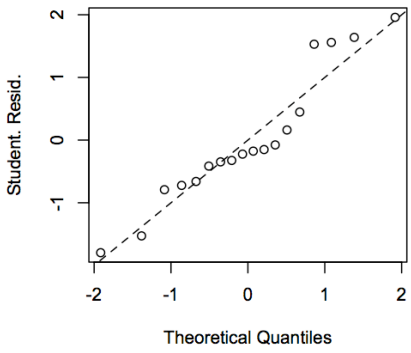

Student. Resid.
Predicted y


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles


y
1/x


Predicted y
y


Student. Resid.
Predicted y


Normal Q-Q Plot
Student. Resid.
Theoretical Quantiles

(2a.) We are given $x^* = 6.0$ and asked to find $y^* = \hat\beta_0 + \hat\beta_1/x^*$ from the second model. Reading the second coefficients table:

$$\widehat{y^*} = \hat\beta_0 + \frac{\hat\beta_1}{x^*} = -5.366 + \frac{85.439}{6.0} = 8.874.$$

(b.) The second is the better model. First, the coefficient of determination $R^2 = .8937$ in the second model whereas it is $R^2 = .6279$ in the first. Thus more of the variation is accounted by Model 2 than by Model 1. Second, looking at the scatterplots (x_i, y_i) in the first panel vs. the scatterplot of $(1/x_i, y_i)$ in the fifth panel, we see that the points are nonlinearly bowed and fall along the hyperbolic curve in the first panel much more than they do in the fifth panel. Although there is scatter, the points seem to be aligning along the dotted line better in the fifth panel. Third, looking at the second and sixth panels, where the predicted values are plotted against the observed ones $(y_i, \hat{y_i})$, the points are bowed in the second panel, whereas they fall nicely along the $\hat{y} = y$ line in the sixth. Thus the second panel indicates that a nonlinear relation is going on whereas it is nicely linear in the sixth. Fourth, looking at the third and seventh panels, where studentized residuals are plotted against predicted values, we see a highly nonlinear pattern for Model 1 whereas the standard deviations seem to uniform without obvious problems in Model 2. The third panel has a "U"-shaped distribution and is highly heteroskedastic: the variance seems to invrease with $\hat{y}$. Finally, the fourth and eighth panels display the PP-plots for the studentized residuals for both models. The fourth panel shows an upward bow whereas the eighth is ragged but not inconsistent with normality. For so few data points, these differences are not especially compelling although the residuals seem to be more normal for Model 2. We have given five features about the fits that indicate that that the second is both the better model and it satisfies the assumptions better.

3. In the multiple regression model we assume $Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \epsilon_i$ for $i = 1, \ldots, n$ where IID $\epsilon_i \sim N(0, \sigma^2)$. Let $x^* = [1, x_1^*, \ldots, x_k^*]$ be a particular value of the independent variables amid the x_i's and $y^* = E(Y|x^*)$ be the expected mean when $x = x^*$.

   (a) Find the estimator $\widehat{y^*}$ and show that it is a linear function of the observed y_i's. [Hint: Find a formula for $\widehat{y^*}$ in terms of x^*, which can be wiewed as a $1 \times p$ matrix ($p = k+1$), the $n \times p$ design matrix $X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \ldots & x_{kn} \end{pmatrix}$, and the $n \times 1$ matrix y.]

   (b) Find the standard error $s(\widehat{y*})$ in terms of s, x^*, X.

(a.) Let X' denote the transpose. The predicted value is $\widehat{y^*} = \hat\beta_0 + \hat\beta_1 x_1^* + \cdots + \hat\beta_k x_k^* = x^* \hat\beta$. The normal equations for the estimates of the coefficients $X'X\hat\beta = X'y$ so

$$\hat\beta = (X'X)^{-1}X'y \qquad \text{and} \qquad \widehat{y^*} = x^*\hat\beta = x^*(X'X)^{-1}X'y = Ay.$$

Since X and x^* are assumed fixed, this shows that $\widehat{y^*}$ is a linear function of y.
(b.) Use the fact that ϵ_i are IID $N(0, \sigma^2)$ so we have $\operatorname{Cov}(\epsilon) = \sigma^2 I$. This implies that $\operatorname{Cov}(y) = \operatorname{Cov}(X\beta + \epsilon) = \sigma^2 I$ also since we've just added a constant. Observe that $\widehat{y^*} = Ay$ is one dimensional so that $\operatorname{V}(\widehat{y^*}) = \operatorname{Cov}(\widehat{y^*})$. Using the formula $\operatorname{Cov}(Ay) = A \operatorname{Cov}(y) A'$,

$$\begin{aligned}\operatorname{V}(\widehat{y^*}) &= \operatorname{Cov}(Ay) = A \operatorname{Cov}(y) A' = \sigma^2 AIA' = \sigma^2 x^*(X'X)^{-1}X' \left(x^*(X'X)^{-1}X'\right)' \\ &= \sigma^2 x^*(X'X)^{-1}X'X \left((X'X)^{-1}\right)' (x^*)' = \sigma^2 x^* \left((X'X)'\right)^{-1} (x^*)' = \sigma^2 x^*(X'X)^{-1}(x^*)'\end{aligned}$$

Substituting the estimator $s^2 = \widehat{\sigma^2} = MSE$, it follows that the standard error

$$\operatorname{s}(\widehat{y^*}) = \operatorname{s} \sqrt{x^*(X'X)^{-1}(x^*)'}.$$

4. A study "...Reaction of Formaldehyde with Cotton Cellulose," (*Textile Research J.*, 1984) of the effect of formaldehyde on fabric measured y durable press rating, a quantitative measure of wrinkle resistance, x_1 HCHO (formaldehyde) concentration, x_2 catalyst ratio, x_3 curing temperature and x_4 curing time.

x1	x2	x3	x4	y	x1	x2	x3	x4	y	x1	x2	x3	x4	y
8	4	100	1	1.4	4	10	160	5	4.6	10	1	180	1	2.6
2	4	180	7	2.2	4	13	100	7	4.3	2	13	140	1	3.1
7	4	180	1	4.6	10	10	120	7	4.9	6	13	180	7	4.7
10	7	120	5	4.9	5	4	100	1	1.7	7	1	120	7	2.5
7	4	180	5	4.6	8	13	140	1	4.6	5	13	140	1	4.5
7	7	180	1	4.7	10	1	180	1	2.6	8	1	160	7	2.1
7	13	140	1	4.6	2	13	140	1	3.1	4	1	180	7	1.8
5	4	160	7	4.5	6	13	180	7	4.7	6	1	160	1	1.5
4	7	140	3	4.8	7	1	120	7	2.5	4	1	100	1	1.3
5	1	100	7	1.4	5	13	140	1	4.5	7	10	100	7	4.6

Here is (partial) **SAS** output for two models fitted to the data, the reduced Model$(y|x1, x2)$ and the full linear Model$(y|x1, x2, x3, x4)$. Use the partial F-test at the $\alpha = .05$ level to determine whether the data strongly suggests that the predictors x_3 and x_4 should be included in the full model. State the null and alternative hypotheses, the test statistic, the rejection region and your conclusions.

Model: Reduced Model
Dependent Variable: y durable press rating

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	33.91947	16.95973	19.95	<.0001
Error	27	22.95253	0.85009		
Corrected Total	29	56.87200			

Root MSE	0.92201	R-Square	0.5964
Dependent Mean	3.56000	Adj R-Sq	0.5665
Coeff Var	25.89902		

Model: Full Model
Dependent Variable: y durable press rating

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	39.37694	9.84423	14.07	<.0001
Error	25	17.49506	0.69980		
Corrected Total	29	56.87200			

Root MSE	0.83654	R-Square	0.6924
Dependent Mean	3.56000	Adj R-Sq	0.6432
Coeff Var	23.49837		

(4.) The null and alternative hypotheses are

$$\mathcal{H}_0 : \beta_3 = \beta_4 = 0 \text{ in the full model } y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_4 x_{4i} + \epsilon_i.$$
$$\mathcal{H}_1 : \beta_3 \neq 0 \text{ or } \beta_4 \neq 0 \text{ in the full model.}$$

The test statistic is

$$F = \frac{\dfrac{SSE(r) - SSE(f)}{DF(r) - DF(f)}}{\dfrac{SSE(f)}{DF(f)}}$$

where $SSE(f)$ and $DF(f)$ are the sum squared residuals and residual degree of freedom for the full model and $SSE(r)$ and $DF(r)$ for the reduced models. If $\mathcal{H}_0$ holds, F is distributed like an $f_{DF(r)-DF(f),DF(f)}$ variable. If $\mathcal{H}_0$ holds, $\mathrm{E}(SSE(r) - SSE(f)) = (DF(r) - DF(f))\sigma^2$ but it is larger when $\mathcal{H}_0$ fails, thus the null hypothesis is rejected if $F > f_{DF(r)-DF(f),DF(f)}(\alpha)$.

Substituting SSE's from the ANOVA tables,

$$F = \frac{\dfrac{SSE(r) - SSE(f)}{DF(r) - DF(f)}}{\dfrac{SSE(f)}{DF(f)}} = \frac{\dfrac{22.95253 - 17.49506}{27 - 25}}{\dfrac{17.49506}{25}} = 3.899294.$$

Now, $f_{DF(r)-DF(f),DF(f)}(\alpha) = f_{2,25}(.05) = 3.39 < F$ so we reject the null hypothesis: the data suggests strongly that β_3 and β_4 are not both zero in the full model.

5. The article "Optimum Design of an A-Pillar Trim…," (Proc. Inst. Mechanical Engineers, 2001) compared different types of automobile head pillars and the protection they give in collisions. Three types of spacing arrangements were tested and the head injury criterion (HIC) was measured for each in nine replications. Partial MINITAB output is given.

```
One-way Analysis of Variance

Analysis of Variance for HIC
Source     DF           SS           MS            F            P
Spacing     2     509646.6      25473.3        5.071        0.015
Error      24     120550.9       5023.0
Total      26     171497.4

Level       N         Mean
A           9       930.87
B           9       873.14
C           9       979.41
```

Does the data strongly indicate that the "B" spacing results in a smaller HIC than the "C" spacing? State the null and alternative hypotheses, the test statistic and the result of your test.

In the one-way fixed effects ANOVA model, the dependent variable y (HIC) satisfies $y_{ij} = \mu_i + \epsilon_{ij}$ where IID $\epsilon_{ij} \sim N(0, \sigma^2)$ for $i = 1, \ldots, I$ the treatments (factors) and $j = 1, \ldots, J$ the replications for the ith factor. Let $n = IJ$. We are testing a contrast, the difference of two means. The null and alternative hypotheses are

$$\mathcal{H}_0 : \mu_2 - \mu_3 = 0;$$
$$\mathcal{H}_1 : \mu_2 - \mu_3 < 0.$$

The test statistic is

$$T = \frac{\overline{Y}_{2\cdot} - \overline{Y}_{3\cdot}}{s(\overline{Y}_{2\cdot} - \overline{Y}_{3\cdot})}$$

which is distributed according to t_{n-I}. We reject $\mathcal{H}_0$ if $T < -t_{n-I}(\alpha)$ since this is a one-sided test. In our case, $I = 3$, $J = 9$, $n = IJ = 27$, $n - I = 24$ and $t_{n-I}(\alpha) = t_{24}(.05) = 1.711$ and

$$s(\overline{Y}_{2\cdot} - \overline{Y}_{3\cdot}) = \sqrt{\frac{2 \cdot MSE}{J}} = \sqrt{\frac{2(5023.0)}{9}} = 33.40991.$$

Thus

$$T = \frac{\overline{Y}_{2\cdot} - \overline{Y}_{3\cdot}}{s(\overline{Y}_{2\cdot} - \overline{Y}_{3\cdot})} = \frac{873.14 - 979.41}{33.40991} = -3.180793.$$

Thus we reject the null hypothesis: the data strongly indicates that $\mu_2 < \mu_3$.