Math 3080 § 1. **First Midterm Sample Rroblems** Name: Sample
Treibergs $\mathscr{A}\mathscr{T}$ February 3, 2010

Problems 1-2 taken from Prof. Roberts' Math 3080 Exams given Spring 2004.

(1.) *Suppose y = amount of residual chlorine in the pool (ppm) and x = hours of cleaning it. It was decided that the model $y = Ae^{Bx}$ would best explain the relationship between x and y. A simlpe linear regression of $\ln y$ on x gave the following results*

Regression equation:	$\ln(\text{chlorine}) = 0.558 - 0.0239(\text{hours})$
	S $= 0.05635$ R-sq $= .759$ R-sq(adjusted) $= 0.699$
Summary Statistics for:	
X (hours):	n$= 6$ Mean$=7.00$ StDev $= 1.53$

i) *What would you use for A, B in the model $Y = Ae^{Bx}$?*

Taking logs we obtain the intrinsically linear model $\ln Y = \ln A + Bx$ so that $A = \exp(.558) = 1.747$ and $B = -0.0239$.

ii) *Based on the model, give an interval estimate (95%) for the mean amount of chlorine (ppm) in the pool 15 hours after the next cleaning. Show your work.*

The estimate for the mean amount of chlorine after $x^* = 15$ hours is $\ln(y)^* = 0.558 - 0.0239x^* = 0.558 - 0.0239(15) = .1995$ so that $y^* = \exp(.1995) = 1.2208$. We need to find S_{xx} which is given by

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = (n-1)(\text{Std. Dev.})^2 = 5 \cdot (1.53)^2 = 11.70.$$

The 95% or $\alpha = .05$ confidence interval for $E(\ln(Y)|X = 15)$ is given by $\ln(y)^* \pm t_{\alpha/2,n-2} s_{\ln(y)^*}$ where $t_{.025,4} = 2.776$ and

$$s_{\ln(y)^*} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} = 0.05635\sqrt{\frac{1}{6} + \frac{(15-7)^2}{11.70}} = .1338$$

It follows that the CI for $E(\ln(Y)^*)$ is $1.2208 \pm 2.776 \cdot 0.1338$ or $(.849, 1.592)$. Exponentiating, the corresponding interval for Y is $(2.237, 4.914)$.

(2.) *Data on the gain of reading speed (y words/min.) and the number of weeks in a speed reading program (x) was recorded for 10 students selected at random from the program. A statistical software program indicated that the simple linear regression model for y was statistically significant, and provided the following information. Use this to answer the following questions.*

Regression equation:	sp.gain $= 2.64 + 11.8$ weeks
	$\text{S}^2 = \text{MSE} = 115.6$
Summary Statistics for	
No of Wks (x):	mean $= 6.1$; standard deviation $= 2.807$; min $= 2$; max $= 11$
Gain in speed (y):	mean $= 69.1$; standard deviation $= 33.3$; min $= 21$; max $= 130$

i) *Give a 95% confidence interval estimate for the mean gain in speed for students who have been in the program for 9 weeks.*

The expected value at $x^* = 9$ weeks is $y^* = 2.64 + 11.8 \cdot 9 = 108.84$. We need to find S_{xx} which is given by

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = (n-1)(\text{Std. Dev.})^2 = 9 \cdot (2.807)^2 = 70.91.$$

The 95% or $\alpha = .05$ confidence interval for $E(Y|X = 9)$ is given by $y^* \pm t_{\alpha/2,n-2}s_{y^*}$ where $t_{.025,8} = 2.306$ and

$$s_{y^*} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} = \sqrt{115.6}\sqrt{\frac{1}{10} + \frac{(9-6.1)^2}{70.91}} = 5.027$$

It follows that the CI for $E(Y^*)$ is $108.84 \pm 2.306 \cdot 5.027$ or $(97.25, 120.43)$.

ii) *What was the R-sq value for this model and what does it measure in the practical context of the problem?*

Using the fact that $\hat{\beta}_1 = S_{xy}/S_{xx}$, $S_{yy} = (n-1)(\text{St.Dev.}y)^2 = 9(33.3)^2 = 9980.01$ and S_{xx} from above, we find that

$$R = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \hat{\beta}_1\sqrt{\frac{S_{xx}}{S_{yy}}} = 11.8\sqrt{\frac{70.91}{9980.01}} = .995.$$

$R^2 = .990$ is the coefficient of determination, which says under the hypotheses of the linear regression model, the proportion of the observed speed gain in reading accounted for by the number of weeks in the program through the simple linear model is 99.0%.

Problems 3-4 taken from my Math 3080 Exam given March 30, 2005.

(3) *A random sample of 18 measurements of the amount of a chemical compound y (g) dissolved in 100g of water at a temperature x (C°) was taken. The simple linear regression model for y was statistically significant. Use the following partial MacAnova printout to answer the following questions*

```
.
.  Model used is y=x
.                    Coef        StdErr            t       P-Value
.  CONSTANT        5.8254         1.075       5.4191    5.6786e-05
.  x              0.56762       0.02367        23.98       < 1e-08
.
.                     DF            SS            MS             F      P-value
.  CONSTANT            1         13230         13230    1999.05025      < 1e-08
.  x                   1        3805.9        3805.9     575.05888      < 1e-08
.  ERROR1             16        105.89        6.6183
```

(i.) [8] Does the data suggest with 95% confidence that $\beta_1 > 0.5$?
(ii.)[8] Find the R^2 value.

To test the null hypothesis $\mathcal{H}_0 : \beta_1 = 0.5$ vs $\mathcal{H}_a : \beta_1 > 0.5$ we compute the t-statistic, and accept the alternate hypothesis provided that $t > t_{\alpha,n-2} = t_{.05,19} = 1.729$. The printout gives both $\hat{\beta}_1 = 0.56762$ and $s_{\hat{\beta}_1} = 0.02367$. Computing t, we have

$$t = \frac{\hat{\beta}_1 - 0.5}{s_{\hat{\beta}_1}} = \frac{0.56762 - 0.5}{0.02367} = 2.857.$$

Thus the alternate hypothesis is accepted, with 95% confidence, $\beta_1 > 0.5$.

The printout tells us that $SSE = 105.89$ and $SSR = 3805.9$. Thus the coefficient of determination is given by

$$r^2 = \frac{SSR}{SST} = \frac{SSR}{SSE + SSR} = \frac{3805.9}{105.89 + 3805.9} = .97293.$$

(4) *Suppose n points (x_i, y_i) are given such that not all the x_i's are equal. Let $y = \hat{\beta}_0 + \hat{\beta}_1 x$ be the ordinary least squares regression line through the points, let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the predicted values and let $\hat{e}_i = y_i - \hat{y}_i$ be the residuals. Show that*

$$\sum_{i=1}^{n} \hat{e}_i = 0.$$

The fact that the x_i's are not all equal means that $S_{xx} > 0$. Using $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ the formulas for the coefficients $\hat{\beta}_1 = S_{xy}/S_{xx}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$,

$$\begin{aligned}\sum_{i=1}^n \hat{e}_i = \sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = \sum_{i=1}^n \left(y_i - \left(\bar{y} - \hat{\beta}_1\bar{x}\right) - \hat{\beta}_1 x_i\right)\\ = \sum_{i=1}^n (y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x}) = n\bar{y} - n\bar{y} - \hat{\beta}_1 (n\bar{x} - n\bar{x}) = 0.\end{aligned}$$

More Probblems.

(5) Consider the regression model with no intercept given by $y_i = \beta x_i + \varepsilon_i$ where $\beta \in \mathbf{R}$ is unknown. For fixed x_i's, suppose that we observe the independent values $y_1, y_2, \ldots, y_n$. Determine the least squares estimator $\hat{\beta}$ for β. Show that your estimator is unbiased. Find the standard error of your estimator. Show that MSE is an unbiased estimator for σ^2. Assume that the random sample $\varepsilon_i \sim N(0, \sigma)$ is taken from a normal distribution with fixed a variance.

The $\hat{\beta}$ is chosen to have least sum of squares of residuals. Thus we need to minimize the function

$$L(b) = \sum_{i=1}^n (y_i - bx_i)^2 .$$

This quadratic function is minimized where the derivative vanishes, or

$$0 = \frac{\partial L}{\partial b} = -\sum_{i=1}^n 2(y_i - bx_i)x_i = 2\sum_{i=1}^n x_i y_i - 2\beta\sum_{i=1}^n x_i^2$$

so that assuming at least one of the x_i's are not zero,

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{Q}$$

where $Q = \sum_{j=1}^n x_j^2$. To see that $\hat{\beta}$ is unbiased estimator for β. We compute the expectation for fixed x_i's using the fact that the expectation of a linear combination is the combination of expectations.

$$E(\hat{\beta}) = \sum_{i=1}^n \frac{x_i}{Q} E(y_i) = \sum_{i=1}^n \frac{x_i}{Q}\beta x_i = \frac{Q}{Q}\beta = \beta.$$

To compute the variance of the statistic, using the independence of the y_i's,

$$V(\hat{\beta}) = \sum_{i=1}^n \frac{x_i^2}{Q^2} V(y_i) = \frac{\sigma^2}{Q}.$$

Computing the expectation of $SSE = \sum_{i=1}^n (y_i - \hat{\beta}x_i)^2$, use for any rv, $E(Z^2) = V(Z) + E^2(Z)$,

$$\begin{aligned}E(SSE) =& \sum_{i=1}^n E\left((y_i - \hat{\beta}x_i)^2\right) = \sum_{i=1}^n V\left(y_i - \hat{\beta}x_i\right) + \sum_{i=1}^n \left(E\left(y_i - \hat{\beta}x_i\right)\right)^2\\ =& \sum_{i=1}^n V\left(y_i - \frac{\sum_{j=1}^n x_j y_j}{Q} x_i\right) + \sum_{i=1}^n (\beta x_i - \beta x_i)^2 = \sum_{i=1}^n V\left(\left(1 - \frac{x_i^2}{Q}\right) y_i - \sum_{j\neq i} \frac{x_i x_j}{Q} y_j\right)\\ =& \sum_{i=1}^n \left(\left(1 - \frac{x_i^2}{Q}\right)^2 V(y_i) + \sum_{j\neq i}\left(\frac{x_i x_j}{Q}\right)^2 V(y_j)\right) = \sum_{i=1}^n \left(\left(1 - \frac{2x_i^2}{Q}\right)\sigma^2 + \sum_{j=1}^n \left(\frac{x_i x_j}{Q}\right)^2 \sigma^2\right)\\ =& \sigma^2 \sum_{i=1}^n \left(\left(1 - \frac{2x_i^2}{Q}\right) + \frac{x_i^2}{Q}\right) = (n-1)\sigma^2.\end{aligned}$$

(6) *A windmill is used to generate direct current. Data are collected on 45 different days to determine the relationship between wind speed x mi/h and current y kA. Compute the least squares line for predicting Model I: y from x and the least squares line from predicting Moldel II: y from $\ln x$. Which of these two models fits best?*

Wind Speed	Curr.
4.2	1.9
1.4	0.7
6.6	2.2
4.7	2.0
2.6	1.1
5.8	2.6
7.7	2.6

Wind Speed	Curr.
1.8	0.3
5.8	2.3
7.3	2.6
7.1	2.7
6.4	2.4
4.6	2.2
6.6	2.9

Wind Speed	Curr.
1.6	1.1
2.3	1.5
4.2	1.5
3.7	2.1
5.9	2.2
6.0	2.6
6.9	2.6

Wind Speed	Curr.
10.7	3.2
5.3	2.3
5.1	1.9
4.9	2.3
8.3	3.1
7.1	2.3

Wind Speed	Curr.
9.2	2.9
4.4	1.8
8.0	2.6
10.5	3.0
5.1	2.1
5.8	2.5

Wind Speed	Curr.
2.6	1.4
7.7	2.8
6.1	2.4
5.5	2.2
4.7	2.3
4.0	2.0

Wind Speed	Curr.
2.3	1.2
11.9	3.0
8.6	2.5
5.6	2.1
4.2	1.7
6.2	2.3

For Model I, the regression line is $y = 0.833 + 0.235x$. For Model II, the regression line is $y = 0.199 + 1.207\ln(x)$.

Compare scatterplots for (x_i, y_i) and $(\ln(x_i), y_i)$. The straight lines are the least squares regression lines. The curved line in the first is the fitted line from Model II, $(x, 0.199 + 1.207\ln(x))$. the curved line in the second is the fitted line from Model I: $(\ln(x), 0.833 + 0.235x)$. Note that the Model II seems to capture the curvature of Model II. The transformation of variables seems to straighten out the cloud of points.

Looking at the ANOVA tables from MacAnova, we see that $r^2 = 0.79242$ for Model I and $r^2 = 0.87899$ for Model II, which means that the transformed points are more linear in Model II. Model II does the better job fitting the line.

```
Model used is Current=WindSp
                 Coef       StdErr            t
CONSTANT      0.83325      0.11355        7.338
WindSp        0.23542     0.018375       12.812
 N: 45, MSE: 0.084662, DF: 43, R^2:  0.79242
Regression F(1,43):  164.15, Durbin-Watson:  1.8832
               DF           SS           MS            F      P-value
CONSTANT        1       213.42       213.42   2520.86932      < 1e-08
WindSp          1       13.897       13.897    164.15015      < 1e-08
ERROR1         43       3.6405     0.084662


 Model used is Current = log(WindSp)
                     Coef       StdErr            t
CONSTANT          0.19878      0.11677       1.7023
log(WindSp)        1.2066     0.068272       17.673
 N: 45, MSE: 0.049354, DF: 43, R^2:  0.87899
Regression F(1,43):  312.34, Durbin-Watson:  1.9299
               DF           SS           MS            F      P-value
CONSTANT        1       213.42       213.42   4324.27902      < 1e-08
log(WindSp)     1       15.416       15.416    312.34371      < 1e-08
ERROR1         43       2.1222     0.049354
```

(7) *A Journal Marketing Research from 1970 study of gasoline pricing reported the following data on $n = 441$ stations. At the 0.01 level, does the data strongly suggest that the facility conditions and pricing policy are not independent? State the null and alternative hypotheses, the test statistic and the rejection region.*

		Observed Pricing Policy		
		Aggressive	Neutral	Nonaggressive
Condition	Substandard	24	15	17
	Standard	52	73	80
	Modern	58	86	36

Math 3080 § 1. **Sample First Midterm Exam** February 3, 2010

This is a χ^2 test of independence. Let R_1 R_2 and R_3 denote the event that the condition is substandard, standard or modern, resp. Let C_1, C_2 and C_3 denote the event that the pricing policy is aggressive, neutral or nonaggressive, resp. The null and alternative hypotheses are

$\mathcal{H}_0 : P(R_i \cap C_j) = P(R_i)P(C_j)$ for all $i, j = 1, 2, 3$. (*i.e.,* condition and pricing are independent.)
$\mathcal{H}_1 : \mathcal{H}_0$ is not true, *i.e.,* $P(R_i \cap C_j) \neq P(R_i)P(C_j)$ for some i, j.

The test statistic is χ^2 as in (†), and the null hypothesis is rejected if $\chi^2 > \chi^2_{\mathrm{df}}(\alpha)$ at the α level of significance.

Summing over rows $y_{\cdot,j}$ and over columns $y_{i,\cdot}$, we can get the table of expected pricing from $e_{ij} = y_{\cdot,j}y_{i,\cdot}/n$. For example, $e_{23} = 133 \cdot 205/441 = 61.83$.

Observed Totals

24	15	17	56
52	73	80	205
52	73	80	180
134	174	133	441

Expected Totals

17.02	22.10	16.89	56
62.29	80.88	61.83	205
54.69	71.02	54.29	180
134	174	133	441

Thus, the test statistic

(†) $$\chi^2 = \sum_{i,j} \frac{(y_{ij} - e_{ij})^2}{e_{ij}} = \frac{(24 - 17.02)^2}{17.02} + \cdots + \frac{(36 - 54.29)^2}{54.29} = 22.47$$

The critical value for df $= (r-1)(c-1) = 2 \cdot 2 = 4$ degrees of freedom is $\chi^2_4(0.01) = 13.277$. Since χ^2 is greater, $\mathcal{H}_0$ is rejected.