Math 3080 § 1.
Treibergs

First Midterm Exam

Name: Solutions
February 10, 2010

1. The article "Flexible Pavement Evaluation..." (*Transportation Eng. J.*, 1977) used simple regression to study the relationship between pavement deflection and surface temperature at various locations on a state highway. Let x be temperature (F°) and y be the deflection adjustment factor. Fill in the ANOVA table. [In each box the correct number is worth one point and the explanation or formula is worth one point.]

$$n = 10, \qquad \sum x_i = 403, \qquad \sum y_i = 570,$$
$$\sum x_i^2 = 18319, \qquad \sum x_i y_i = 23922, \qquad \sum y_i^2 = 33566.$$

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model				
Error				
Total				

$R-$Square	

First compute the sum of squares (denote $\sum_{i=1}^n$ by $\sum$)

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2 = 18319. - \frac{1}{10}(403)^2 = 2078.1$$
$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right) = 23922 - \frac{1}{10}(403)(570) = 951.0$$
$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2 = \sum (y_i - \bar{y})^2 = 33566 - \frac{1}{10}(570)^2 = 1076.0$$

Then the ANOVA table is computed according to the formulas

$$\text{Model d.f.} = 1$$
$$\text{Error d.f.} = n - 2 = 8$$
$$\text{Total d.f.} = n - 1 = 9$$
$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} = \frac{(951)^2}{(2078.1)(1076)} = .404466279$$

$$
\begin{aligned}
SST &= S_{yy} = 1076 \\
SSR &= \frac{S_{xy}^2}{S_{xx}} \; (= R^2 \cdot SST) = \frac{(951)^2}{2078.1} = 435.2057168 \\
SSE &= SST - SSR = 1076.0 - 435.2 = 640.7942832 \\
MSR &= \frac{SSR}{\text{Model d.f.}} = \frac{435.2057168}{1} = 435.2057168 \\
MSE &= \frac{SSE}{\text{Error d.f.}} = \frac{640.7942832}{8} = 80.0992855 \\
F &= \frac{MSR}{MSE} = \frac{435.2057168}{80.0992855} = 5.433328326.
\end{aligned}
$$

The ANOVA table consists of

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	Model d.f.	SSR	MSR	F
Error	Error d.f.	SSE	MSE	
Total	Total d.f.	SST		
$R-$Square	R^2			

Filling in the values

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	1	435.2	435.2	5.4333
Error	8	640.8	80.10	
Total	9	1076.0		
$R-$Square	.4045			

2. The study "Susceptability of Mice to Audiogenic Seizure..." (*Science*, 1976) reports on different injection treatments on the frequencies of seizures. What is the expected count of mice in the study that were treated with Sham that exhibited Wild Running? Does the data suggest that the true percentages in the different response categories depend on the nature of the injection treatment? State and test the appropriate hypothesis at the $\alpha = .005$ level using the (partial) SAS output.

```
   The FREQ Procedure:

                 Table of treatment by response

treatment        response

Frequency       |
Percent         |
Row Pct         |
Col Pct         |No      |Wild    |Clonic  |Tonic   |  Total
                |Response|Running |Seizure |Seizure |
----------------+--------+--------+--------+--------+
Theinylalanine  |     21 |      7 |     24 |     44 |     96
                |   4.96 |   1.65 |   5.67 |  10.40 |  22.70
                |  21.88 |   7.29 |  25.00 |  45.83 |
                |  19.81 |  15.91 |  25.26 |  24.72 |
----------------+--------+--------+--------+--------+
Solvent         |     15 |     14 |     20 |     54 |    103
                |   3.55 |   3.31 |   4.73 |  12.77 |  24.35
                |  14.56 |  13.59 |  19.42 |  52.43 |
                |  14.15 |  31.82 |  21.05 |  30.34 |
----------------+--------+--------+--------+--------+
Sham            |     23 |     10 |     23 |     48 |    104
                |   5.44 |   2.36 |   5.44 |  11.35 |  24.59
                |  22.12 |   9.62 |  22.12 |  46.15 |
                |  21.70 |  22.73 |  24.21 |  26.97 |
----------------+--------+--------+--------+--------+
Unhandled       |     47 |     13 |     28 |     32 |    120
                |  11.11 |   3.07 |   6.62 |   7.57 |  28.37
                |  39.17 |  10.83 |  23.33 |  26.67 |
                |  44.34 |  29.55 |  29.47 |  17.98 |
----------------+--------+--------+--------+--------+
Total                106       44       95      178      423
                   25.06    10.40    22.46    42.08   100.00

       Statistics for Table of treatment by response
   Statistic                      DF       Value      Prob
   ------------------------------------------------------
   Chi-Square                      9     27.6642    0.0011
                      Sample Size = 423
```

The expected count in the (Sham, Wild Running) cell is

$$e_{3,2} = n_{3,\bullet} \cdot \hat{p}_2 = n_{3,\bullet} \cdot \frac{n_{\bullet,2}}{n_{\bullet,\bullet}} = \frac{104 \cdot 44}{423} = 10.818.$$

This is a χ^2-test of homogeneity. Let p_{ij} denote the probability that the a mouse receiving the i-th treatment will exhibit the jth response. The null hypothesis is that the response does not depend on the treatment, or,

$$\mathcal{H}_0 : p_{1j} = p_{2j} = p_{3j} = p_{4j} = p_j \text{ for all } j = 1, 2, 3, 4.$$
$$\mathcal{H}_1 : \mathcal{H}_0 \text{ is false: } p_{ij} \neq p_{i'j} \text{ for some } i, i', j \in \{1, 2, 3, 4\} \text{ where } i \neq i'.$$

The output computes the statistic

$$\chi^2 = \sum_{i=1}^{4} \sum_{j=1}^{4} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = 27.6642.$$

Since all expected cell sizes $e_{ij} \geq 5$ (since SAS did not give an expected count below 5 warning), the statistic has approximately the χ^2 distribution with $(r-1)(c-1) = 3 \cdot 3 = 9$ degrees of freedom. According to the output, the P-value is $.0011 < \alpha = .005$ so we reject the null hypothesis: there is strong evidence that the response depends on the treatment.

3. The study of the sterility of a fruit fly "Hybrid Dysgenesis..." (*Genetics*, 1979) proposed that the number of ovaries that develop is a binomial random variable with density

$$p(x) = \binom{2}{x} p^x (1-p)^{2-x}, \qquad \text{for } x = 0, 1, 2$$

for some $0 < p < 1$. Test whether the data is consistent with this model.

[Hint: the MLE turns out to be $\hat{p} = \dfrac{n_1 + 2n_2}{2(n_0 + n_1 + n_2)}$ which is $\hat{p} = .0843$ for these numbers.]

$x =$ No. Ovaries Developed:	0	1	2
$n_x =$ Observed Count:	1212	118	58

We use a χ^2-goodness of fit test where the cell probabilities are not completely specified. Using the MLE for p gives the estimate of cell probabilities and cell frequencies $e_x = p(x)n$ where $\hat{p}$ is used to compute $p(x)$. Thus

$x =$ No. Ovaries Developed:	0	1	2	Total
$n_x =$ Observed Count:	1212	118	58	1388
$p(x) =$ Model Cell Prob.	$(1-p)^2$	$2p(1-p)$	p^2	1
Estimated $p(x)$ (not needed)	.83850649	.15438702	.00710649	1
$e_x = np(x) =$ Expected Cell Count	1163.847008	214.289184	9.863808	1388
$\dfrac{(n_x - e_x)^2}{e_x}$	1.992281	43.266798	234.908561	280.1676

Since we estimated $k = 1$ parameter, the χ^2 statistic has approximately a χ^2 distribution with $c - 1 - k = 3 - 1 - 1 = 1$ degree of freedom for large n. Since the expected cell counts exceed 5, by our rule of thumb, the test is applicabile.

The null hypothesis is

$$\mathcal{H}_0 : P(\ x \text{ ovaries develop}) = p(x) \text{ for } x = 0, 1, 2 \text{ and for some parameter } 0 < p < 1$$

Using the MLE for p gives the χ^2 statistic

$$\chi^2 = \sum_{x=0}^{2} \frac{(n_x - e_x)^2}{e_x} = 280.2.$$

The critical value for one degree of freedom $\chi_1^2(.005) = 7.879$. Since our statistic is greater, we reject the null hypothesis: the data indicates stongly that the binomial distibution does not provide a good model.

4. Consider the simple regression model for $i = 1, \ldots, n$

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

where $\epsilon_i \sim N(0, \sigma^2)$ are IID normal random variables. Let $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the predicted value. Show that

$$\sum_{i=1}^{n} x_i (Y_i - \hat{Y}_i) = 0.$$

Using the formula $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, the predicted values can be rewritten

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{Y} - \hat{\beta}_1 \bar{x} + \beta_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x}).$$

Thus

$$\begin{aligned}
\sum_{i=1}^{n} x_i (Y_i - \hat{Y}_i) &= \sum_{i=1}^{n} \Big\{ x_i (Y_i - \bar{Y}) - x_i \hat{\beta}_1 (x_i - \bar{x}) \Big\} \\
&= \sum_{i=1}^{n} \Big\{ x_i (Y_i - \bar{Y}) - \bar{x}(Y_i - \bar{Y}) - x_i \hat{\beta}_1 (x_i - \bar{x}) + \bar{x} \hat{\beta}_1 (x_i - \bar{x}) \Big\} \\
&= \sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y}) - \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= S_{xy} - \frac{S_{xy}}{S_{xx}} \cdot S_{xx} \\
&= 0.
\end{aligned}$$

We have used the formula $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$ where S_{xx} and S_{xy} are defined as in the solution to Problem 1, and the fact that

$$\sum_{i=1}^{n} \left(Y_i - \bar{Y} \right) = \left(\sum_{i=1}^{n} Y_i \right) - n \cdot \left(\frac{1}{n} \sum_{i=1}^{n} Y_i \right) = 0$$

and similarly $\sum (x_i - \bar{x}) = 0$ so that

$$\sum_{i=1}^{n} \bar{x} \left(Y_i - \bar{Y} \right) = 0 = \sum_{i=1}^{n} \bar{x} \hat{\beta}_1 \left(x_i - \bar{x} \right).$$

5. A study of the strength of titanium welds by Harwig *et. al.*, (*Welding Journal,* 2001), compared the oxygen content in parts per thousand (x_i) to strength in ksi (y_i). The model is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where ϵ_i are IID $N(0, \sigma^2)$ variables. What is the estimate for the expected strength if the oxygen content is 1.70 parts per thousand? Find an $\alpha = .05$ lower one sided confidence interval for β_1. Does the data strongly indicate that $\beta_1 > 10.00$? Formulate the null and alternative hypotheses. Test at the $\alpha = 0.05$ level.

```
R version 2.7.2 (2008-08-25)
Copyright (C) 2008 The R Foundation for Statistical Computing
> mean(OxygenContent); mean(Strength)
[1] 1.519655
[1] 75.49655
> fit <- lm(Strength ~ OxygenContent); summary(fit); anova(fit)
Call:
lm(formula = Strength ~ OxygenContent)
Residuals:
     Min       1Q   Median       3Q      Max
-10.0185  -3.6639  -0.1139   4.4977  12.6515
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     49.780      7.751   6.423    7e-07 ***
OxygenContent   16.923      5.050   3.351  0.00239 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.841 on 27 degrees of freedom
Multiple R-squared: 0.2937,Adjusted R-squared: 0.2676
F-statistic: 11.23 on 1 and 27 DF,  p-value: 0.002391

Analysis of Variance Table
Response: Strength
              Df Sum Sq Mean Sq F value   Pr(>F)
OxygenContent  1 383.09  383.09  11.229 0.002391 **
Residuals     27 921.14   34.12
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The estimate of expected strength when $x^* = 1.70$ is done by evaluating the regression line at the point

$$\widehat{\mathbf{E}(Y^*)} = \hat{\beta}_0 + \hat{\beta}_1 x^* = 49.780 + 16.923 \cdot 1.70 = 78.5.$$

$\hat{\beta}_1$ is normally distributed so the standardization using it's standard error is t-distributed with $n-2$ degrees of freedom. Here $n = 29$. Thus with $\alpha = .05$ confidence, $\mathbf{E}(\hat{\beta}_1) = \beta_1$ lies in the lower one-sided confidence interval (from the output)

$$\big(\hat{\beta}_1 - s(\hat{\beta}_1) t_{n-2}(\alpha), \infty\big) = \big(16.923 - (5.050)(1.703), \infty\big) = \big(8.322, \infty\big).$$

The proposed null and alternative hypotheses are

$$\begin{aligned} \mathcal{H}_0 :& \beta_1 = 10.00; \\ \mathcal{H}_1 :& \beta_1 > 10.00. \end{aligned}$$

10.00 lies in the confidence interval above, so that with $\alpha = .05$ confidence we accept the null hypothesis: this study does not provide strong evidence that $\beta_1 > 10.00$.