STATISTICS 174: APPLIED STATISTICS

MIDTERM EXAM

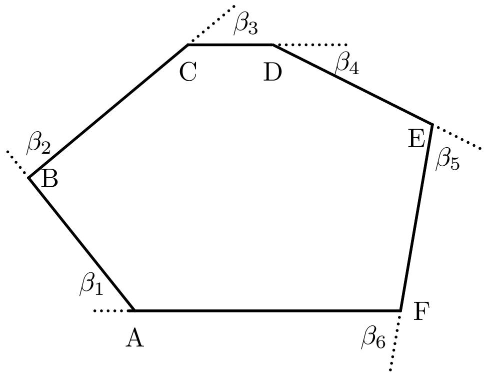
OCTOBER 12, 2004

Time allowed: 75 minutes.

This is an open book exam: all course notes and the text are allowed, and you are expected to use your own calculator. Answers should preferably be written in a blue book.

The exam is expected to be your own work and no consultation during the exam is allowed. You are encouraged to ask the instructor for clarification if the meaning of any parts of the exam are unclear to you.

1. The year is 2011 and the United States is involved in another war in the Middle East. The enemy has established its military headquarters are in a building called the Hexagon, which is known to be in the shape of an irregular six-sided polygon (see figure below). The U.S. military forces need to know the precise shape of this building in order to target their missiles. However they are forced to rely on rather inaccurate surveillance equipment to measure the angles $\beta_1, ..., \beta_6$. From high-school geometry it is known that $\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \beta_6 = 360$ (degrees). However to build some redundancy into the system, they measure all six angles — let $b_1, ..., b_6$ be the actual measurements (which, because of measurement error, need not add up to exactly 360). We do assume that all the measurement errors are independent, of mean 0, and have the same unknown variance σ^2.



   (a) Defining $y_i = b_i$ for $i = 1, 2, 3, 4, 5$, $y_6 = 360 - b_6$, formulate this as a linear model, derive the least squares estimates $\hat{\beta}_1, ..., \hat{\beta}_5$, and state the common variance of these estimates in terms of σ^2. [15 points.]

(b) The U.S. military commander comes across some documents, allegedly stolen from the company that originally constructed the building, that suggest that sides AF and CD are parallel. If this information is correct, it will assist the U.S. forces to plan their assault. Note that under this assumption, we will have $\beta_1 + \beta_2 + \beta_3 = 180$ and $\beta_4 + \beta_5 + \beta_6 = 180$. Using notation similar (but not necessarily identical) to that in part (a), show how the least squares estimates $\hat{\hat{\beta}}_i$ would be constructed in this case, and calculate the (common) variance of these estimates as a function of σ^2. [15 points.]

(c) However, there is some doubt about the authenticity of the documents, so we would like to test the hypothesis that sides AF and CD are indeed parallel. Define the fitted values of the b_i's to be $\hat{b}_i$ under (a) and $\hat{\hat{b}}_i$ under (b). In terms of the quantities $b_i, \hat{b}_i, \hat{\hat{b}}_i$, describe how you would construct an F test of this hypothesis. [10 points.]

2. The data in the attached handout are winning times in the men's 100 meter race in Olympic Games 1900–2004, and in the women's race in 1928–2004. They were analyzed in a recent article in *Nature* ("Momentous sprint at the 2156 Olympics?" by A.J. Tatem, C.A. Guerra, P.M. Atkinson and S.I. Hay; *Nature* **431** p. 525, September 30 2004) who concluded on the basis of linear regression that the women's winning time would eventually be faster than the men's winning time, and they calculated 2156 as the most likely year in which this would happen.

(a) The model

$$y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \epsilon_i \qquad (1)$$

fitted to the men's winning times (for which $\bar{x} = 1954.3333$) yields point estimates $\hat{\beta}_0 = 10.3179$, $\hat{\beta}_1 = -0.01101$, with standard errors .0275 and .000859 respectively; also, the estimated residual standard deviation is 0.1347. The corresponding results for the women's race (for which $\bar{x} = 1968.6667$) are $\hat{\beta}_0 = 11.23$, $\hat{\beta}_1 = -0.01682$, with standard errors .0496 and .002176 respectively, and estimated residual standard deviation is 0.2104. Based on these numbers, calculate a 95% prediction interval for the men's winning time in 2156, and similarly for the women. [15 points.]

(b) Suppose the model (1) is extended to

$$y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \beta_2(x_i - \bar{x})^2 + \beta_3(x_i - \bar{x})^3 + \epsilon_i \qquad (2)$$

The men's results are as follows:

Parameter	Estimate	Standard Error	t value	Pr(> $\|t\|$)
β_0	10.266	.0413	248	0
β_1	–.00988	.00216	–4.58	.00018
β_2	.000048	.000031	1.53	.14
β_3	.00000049	.000001	-0.43	.67

with $s = 0.1320$.

The corresponding results for women are:

Parameter	Estimate	Standard Error	t value	Pr(> $\|t\|$)
β_0	11.135	.0682	163	0
β_1	–.01992	.0053	–3.75	.0021
β_2	.000202	.000102	1.98	.068
β_3	.000004	.000005	0.81	.43

with $s = 0.1989$.

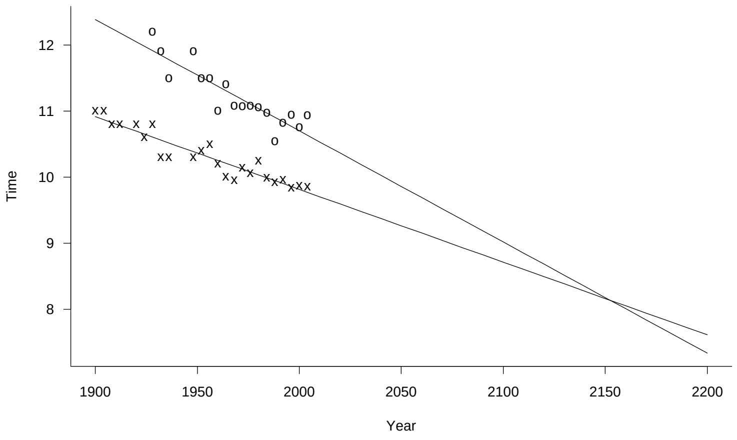
Carry out an F test (separately for the men's and women's races) for the hypothesis that the trend is linear, against the alternative that it follows a cubic regression. What do you conclude? [15 points.]

(c) It is possible that the results are affected by outliers or influential values among the observations. To test this, I have computed the h_i statistics (H), studentized residuals (STUD), DFFITS (DFFI), DFBETAS for the intercept (DFB1) and DFBETAS for the slope (DFB2), for the linear regressions only, for both the men's and women's results. The table appears at the end of this exam. Also included is a plot of fitted lines for both men's and women's races. What do you conclude? [15 points.]

(d) Write a *brief* (10 lines maximum!) critique of the paper overall, indicating whether you agree with the conclusions or if you do not, your main points of disagreement. Feel free to mention other points that have not been covered in (a)–(c). [15 points.]

Table of Diagnostic Statistics

Year	Men H	Men STUD	Men DFFI	Men DFB1	Men DFB2	Women H	Women STUD	Women DFFI	Women DFB1	Women DFB2
1900	0.1618	0.6735	0.2959	0.1502	-0.2550					
1904	0.1448	1.0300	0.4238	0.2273	-0.3576					
1908	0.1290	-0.2166	-0.0834	-0.0474	0.0686					
1912	0.1146	0.1248	0.0449	0.0271	-0.0358					
1920	0.0896	0.8044	0.2524	0.1721	-0.1846					
1924	0.0791	-0.3926	-0.1151	-0.0835	0.0792					
1928	0.0699	1.5238	0.4177	0.3225	-0.2654	0.2324	1.6288	0.8962	0.4382	-0.7818
1932	0.0620	-2.1884	-0.5625	-0.4612	0.3219	0.1993	0.2742	0.1368	0.0722	-0.1162
1936	0.0553	-1.7555	-0.4249	-0.3687	0.2112	0.1697	-1.5156	-0.6851	-0.3920	0.5618
1948	0.0433	-0.6563	-0.1396	-0.1370	0.0271	0.1012	1.7101	0.5739	0.4252	-0.3855
1952	0.0419	0.4197	0.0878	0.0875	-0.0064	0.0853	-0.0499	-0.0152	-0.0123	0.0090
1956	0.0418	1.5697	0.3278	0.3273	0.0170	0.0727	0.2726	0.0763	0.0667	-0.0371
1960	0.0430	-0.4135	-0.0876	-0.0863	-0.0153	0.0636	-2.0139	-0.5248	-0.4905	0.1865
1964	0.0455	-1.6715	-0.3648	-0.3492	-0.1055	0.0579	0.4365	0.1082	0.1060	-0.0217
1968	0.0493	-1.7293	-0.3937	-0.3620	-0.1546	0.0556	-0.7786	-0.1889	-0.1888	0.0055
1972	0.0544	0.1232	0.0295	0.0259	0.0143	0.0567	-0.4964	-0.1217	-0.1205	-0.0176
1976	0.0608	-0.1457	-0.0371	-0.0307	-0.0208	0.0613	-0.1266	-0.0323	-0.0308	-0.0099
1980	0.0685	1.7225	0.4670	0.3643	0.2922	0.0693	0.0985	0.0269	0.0241	0.0120
1984	0.0775	-0.0107	-0.0031	-0.0023	-0.0021	0.0807	-0.0099	-0.0029	-0.0024	-0.0016
1988	0.0878	-0.2083	-0.0646	-0.0445	-0.0468	0.0955	-1.9825	-0.6443	-0.4913	-0.4167
1992	0.0994	0.4347	0.1444	0.0935	0.1101	0.1138	-0.0855	-0.0306	-0.0214	-0.0219
1996	0.1123	-0.1490	-0.0530	-0.0323	-0.0420	0.1354	0.8608	0.3407	0.2182	0.2617
2000	0.1265	0.4261	0.1622	0.0931	0.1328	0.1605	0.2369	0.1036	0.0610	0.0838
2004	0.1421	0.6218	0.2530	0.1370	0.2127	0.1891	1.6322	0.7881	0.4272	0.6622



1. Hexagon problem:

   (a) With $y_1, ..., y_6$ as given and writing $E(y_6) = \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5$, we calculate

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}, \quad X^T X = \begin{pmatrix} 2 & 1 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 1 \\ 1 & 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 1 & 2 \end{pmatrix}, \quad (X^T X)^{-1} = \begin{pmatrix} \frac{5}{6} & -\frac{1}{6} & -\frac{1}{6} & -\frac{1}{6} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{5}{6} & -\frac{1}{6} & -\frac{1}{6} & -\frac{1}{6} \\ -\frac{1}{6} & -\frac{1}{6} & \frac{5}{6} & -\frac{1}{6} & -\frac{1}{6} \\ -\frac{1}{6} & -\frac{1}{6} & -\frac{1}{6} & \frac{5}{6} & -\frac{1}{6} \\ -\frac{1}{6} & -\frac{1}{6} & -\frac{1}{6} & -\frac{1}{6} & \frac{5}{6} \end{pmatrix},$$

   where the expression for $(X^T X)^{-1}$ comes either by guesswork or from the formula for $(aI_n + bJ_n)^{-1}$, with $a = 1,\ b = 1,\ n = 5$. The least squares estimator for β_1 is therefore

$$\frac{5}{6}(y_1 + y_6) - \frac{1}{6}(y_2 + y_3 + y_4 + y_5 + 4y_6) = b_1 - \frac{1}{6}(\sum_1^6 b_i - 360).$$

   The common variance of the $\hat{\beta}_i$'s is $\frac{5}{6}\sigma^2$.

   (b) Considering separately the problems of estimating $(\beta_1, \beta_2, \beta_3)$ and $(\beta_4, \beta_5, \beta_6)$, this is actually the same as the triangle problem discussed in class, e.g. defining $y_1 = b_1,\ y_2 = b_2,\ y_3 = 180 - b_3$, we have optimal estimates $\hat{\beta}_1 = \frac{1}{3}(2y_1 - y_2 + y_3)$ etc., with common variance $\frac{2}{3}\sigma^2$.

   (c) $SSE_1 = \sum(b_i - \hat{b}_i)^2$ with $n - p = 6 - 5 = 1$ degree of freedom. $SSE_0 = \sum(b_i - \hat{\hat{b}}_i)^2$ with 2 degrees of freedom. Therefore the F statistic is

$$F = \frac{SSE_0 - SSE_1}{1} \cdot \frac{1}{SSE_1} = \frac{SSE_0 - SSE_1}{SSE_1}.$$

   The null distribution is $F_{1,1}$ so we would reject at level α if F is larger than the $100(1 - \alpha)$ percentage point of the $F_{1,1}$ distribution.

2. Running times problem:

   (a) For men in 2156, $x - \bar{x} = 201.6667$ so the point predictor is $10.3179 - 201.6667 \times .01101 = 8.098$ with prediction standard error $\sqrt{.1347^2 + .0275^2 + (201.6667 \times .000859)^2} = .2212$ With $t_{22,.975} = 2.074$, a 95% prediction interval is $8.098 \pm 2.074 \times .2212 = (7.639, 8.557)$. Similarly for women, $x - \bar{x} = 187.3333$, the point predictor is 8.079, prediction standard error $\sqrt{.2104^2 + .0496^2 + (187.3333 \times .002176)^2} = .4614$, 95% prediction interval is $8.079 \pm 2.120 \times .4614 = (7.100, 9.057)$.

(b) Men's results: under H_0 (linear regression), $s = .1347$ with $n - p = 24 - 2 = 22$ DF, so $SSE_0 = 22 \times .1347^2 = .39917$. Under H_1, $s = .1320$ with 20 DF, so $SSE_1 = 20 \times .1320^2 = .34848$. The F statistic is

$$F = \frac{SSE_0 - SSE_1}{2} \cdot \frac{20}{SSE_1} = 1.45$$

and this is not rejected at the 10% level (the 90% point of the $F_{2,20}$ distribution is 2.59, from the tables).

Corresponding results for women: $SSE_1 = 16 \times .2104^2 = .70829$, $SSE_0 = 14 \times .1989^2 = .55386$, $F = 1.95$. The 90% point of the $F_{2,15}$ distribution (nearest to true DF) is 2.70, so we do not reject H_0.

(c) Using the rules of thumb given in class, the critical values for H are .1667 (men), .2222 (women), based on $p = 2$, $n = 24$ for men and $n = 18$ for women.

Critical values for studentized residuals (at 5% two-sided level of significance) are 2.080 for men, 2.131 for women, based on t_{21} and t_{15} distributions respectively.

Critical values for DFFITS are .5774 for men, .6667 for women.

Critical values for DFBETAS are .4082 for men, .4714 for women.

For men: no points of high leverage. Possible outlier in 1932. No influential values by DFFITS, though 1932 is close. According to DFBETAS, 1932 is influential on the intercept but not on the mean.

For women: 1928 is a point of high leverage. Possible outliers in 1960 and 1988, though they are within the 95% significance bounds. 1928, 1936 and 2004 are influential according to DFFITS. Several values are influential according to DFBETAS, e.g. 1928, 1936, 2004 for the slope.

(d) Obviously there are many points you could make here but possibilities include: (i) the need to test a wider range of models (e.g. the answer to (b) confirms that a cubic regression is not statistically significant against the null hypothesis of a linear regression, but this does not "prove" that the linear model is correct; (ii) need to take into account outliers and influence (how sensitive are the results to removing one or more data points?); (iii) lack of physical reality of the model, e.g the fact that it would eventually predict negative running times; (iv) the general lack of credibility of any statistical analysis that projects into the far distant future based on a limited period of data. I also felt (though this is a more subjective point) that the article was placing too much emphasis on R^2 as an overall measure of fit, implying that because the linear model had a high R^2, therefore it must be good for prediction — true up to a point, but not sufficient to justify this kind of extrapolation!

Additional Problems

(a) (Continuation of Question 1). The decision is taken to accept the document and proceed according to the estimates in (b). However, at the last moment a second document is discovered, giving a completely different set of plans for the building. *Which document is the forgery?* According to the second document, the values of $\beta_1, ..., \beta_6$ are $\beta_1 = 45$, $\beta_2 = 45$, $\beta_3 = 45$, $\beta_4 = 135$, $\beta_5 = 60$, $\beta_6 = 30$.

Suppose these are the true β's. Suppose also the true value of σ (the standard deviation of a single observation) is 3 degrees. What is the probability that the test in (c) would erroneously lead to the acceptance of the hypothesis that $\beta_1 + \beta_2 + \beta_3 = 180$?

(b) (Continuation of Question 2). Now let's assume that the linear regression (with normal errors, equal variances, etc.) really is the correct model for the data, and the women's rate of improvement is faster than the men's, so that women will indeed eventually overtake men. The paper uses a method we haven't studied in this course, Markov chain Monte Carlo, to conclude that a 95% confidence interval for the first year in which the women's expected time is faster than the men's is (2064, 2788). Suggest a way of doing this, without using Markov chain Monte Carlo, based on the techniques we have learned in the course.

Solutions to Additional Problems

(a) This is essentially asking what is the power of the F test, given the β's quoted. We follow the "substitution rule": the distribution of the F statistic is non-central $F'_{\delta,1,1}$, where $\sigma^2\delta^2$ is $SSE_0 - SSE_1$ when the original data values $b_1, ..., b_6$ are substituted by their values under the alternative hypothesis, i.e. by 45, 45, 45, 135, 60, 30.

SSE_1 is $\sum(b_i - \hat{b}_i)^2$. However, by the formula given for part (a) (or just by common sense), if $\sum_1^6 b_i = 360$, then $\hat{b}_i = b_i$ for all i, so $SSE_1 = 0$. (*Side comment:* it very often happens in applying the substitution rule to determine the power of an F test, that $SSE_1 = 0$.)

To compute SSE_0 we need to calculate the $\hat{\hat{b}}_i$. Applying the answer to part (b), if the b's are as given then $\hat{\hat{b}}_i = 60, 60, 60, 120, 45, 15$ for $i = 1, ..., 6$. Thus $\sum(b_i - \hat{\hat{b}}_i)^2 = 6 \times 15^2 = 1350$. This is $\sigma^2\delta^2$: we were given $\sigma = 3$, and also $\phi = \frac{\delta}{\sqrt{2}}$ so $\phi = \sqrt{\frac{1350}{18}} = 8.66$. But if you apply the S-PLUS command `pearsonhartley(8.66,.05,1,1)` (using the `pearsonhartley` function from the web page) the answer is returned as 0.66. In other words, despite the apparently substantial discrepancy (compared with σ) between $\beta_1 + \beta_2 + \beta_3 = 135$ and the hypothesized value 180, there is still a 34% chance that the F test would incorrectly accept H_0. (*Note:* The values of ϕ and the final power of the test are slightly different from those derived in class, owing to a calculation error on my part.)

(b) There are different ways of doing this (and no unique "right answer", which is one reason I didn't include this in the exam), but here are two possibilities.

(i) Write the men's and women's models as $E(y) = \beta_0 + \beta_1(x - \bar{x}_m)$ and $E(y) = \gamma_0 + \gamma_1(x - \bar{x}_f)$ respectively, where $\bar{x}_m$ and $\bar{x}_f$ are the mean x values for the men's and women's regressions. A 95% confidence interval for the value x at which the two regression lines cross consists of all values for which the hypothesis $H_0 : \beta_0 + \beta_1(x - \bar{x}_m) = \gamma_0 + \gamma_1(x - \bar{x}_f)$ is accepted at level 0.05. We therefore define a test statistic $T = \hat{\beta}_0 + \hat{\beta}_1(x - \bar{x}_m) - \hat{\gamma}_0 - \hat{\gamma}_1(x - \bar{x}_f)$; under H_0 this has mean 0 and variance approximately $\sigma_T^2 = SE(\beta_0)^2 + SE(\beta_1)^2(x - \bar{x}_m)^2 + SE(\gamma_0)^2 + SE(\gamma_1)^2(x - \bar{x}_f)^2$ where SE denotes the standard errors of the respective parameter estimates. Then $\frac{T}{\sigma_T}$ has a standard normal distribution: reject H_0 whenever $|T| > 1.96\sigma_T$. The drawback of this method is that it treats the standard errors as the actual standard deviations of the respective parameter estimates, ignoring that they are themselves estimates based on the residual standard deviations.

(ii) This solution is the same as (i) up to the definition of T. At this point, *if* we assume that the values of σ^2 are the same for both men and women, we can proceed as follows. Let s_m^2 and s_f^2 be the estimated s^2 values for men and women. Recall that these have 22 and 16 degrees of freedom respectively. Therefore, an estimate of the combined s^2 is

$$s^2 = \frac{22s_m^2 + 16s_f^2}{38}.$$

Under this new assumption, the revised standard errors of $\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}_0, \hat{\gamma}_1$ are $\frac{s}{\sqrt{n_m}}, \frac{s}{\sqrt{A}}, \frac{s}{\sqrt{n_f}}, \frac{s}{\sqrt{B}}$ where n_m and n_f are the respective sample sizes (24 and 18), and A and B are the values of $\sum (x_i - \bar{x})^2$ for men and women respectively. Therefore, under this null hypothesis,

$$\frac{T}{s\sqrt{\frac{1}{n_m} + \frac{(x-\bar{x}_m)^2}{A} + \frac{1}{n_f} + \frac{(x-\bar{x}_f)^2}{B}}} \tag{3}$$

has a t_{38} distribution. The resulting test is therefore based on (3).

Note, however, that this second solution doesn't work if we don't assume that σ^2 is the same for both regressions. The situation is similar to the two-sample t test discussed in elementary statistics courses: if both samples are assumed to have the same σ^2, then there is an exact solution leading to a test statistic with a t distribution, but if we don't make that assumption, no exact solution exists (in advanced statistics tests this is known as the Behrens-Fisher problem). In that case there are various possible alternative approaches, one of which is a Bayesian formulation that could be solved by Markov chain Monte Carlo. It's possible that this is what the authors of the *Nature* paper actually did, but they don't say, and I cannot figure it out.