ST505/697R: Fall 2012. Homework 3 Solution.

Throughout I have used exact table values, which can be obtained via SAS or R. If you use approximations from the tables, your answers will be a little different.

1. Neopterin Assay example. This is inverse prediction with $Y_{new} = 3000$. From the regression analysis, $b_0 = 6683.97035$, $b_1 = -1072.92715$, $s^2\{b_0\} = 19817.939942$, $s\{b_0, b_1\} = -4707.512482$, $s^2\{b_1\} = 1276.1361656$, MSE =68670 and $n = 28$. The regression is in terms of log concentration. Running the analysis on this we obtain an estimated log(concentration), an estimated approximate standard error, and an approximate confidence interval given by
$\hat{X}_{new} = (3000 - b_0)/b_1 = 3.4335$
$s_{predX} = 0.2487065$
$\hat{X}_{new} \pm 2.055 * s_{predX} = [2.922, 3.944]$,
where $t(.975, 26) = 2.055$ and $s_{predX} = \sqrt{0.06185}$,

$$s^2\{predX\} = .06185 = \frac{1}{b_1^2}(MSE + s^2\{b_0\} + \hat{X}_{new}^2 s^2\{b_1\} + 2\hat{X}_{new} s\{b_0, b_1\}).$$

The Fieller interval uses

$$\frac{c_{01}}{c_1} \pm \frac{[c_{01}^2 - c_0 c_1]^{1/2}}{c_1},$$

where $c_0 = (Y_{new} - b_0)^2 - t^2(MSE + s^2\{b_0\})$, $c_1 = (b_1)^2 - t^2 s^2\{b_1\}$ and $c_{01} = (Y_{new} - b_0)b_1 + t^2 s\{b_0, b_1\}$, with quantities going into these as given above.

If you calculate out the Fieller interval is $[2.920, 3.945]$, which is essentially the same as the approximate method.

Since the concentration $= e^x$ is a monotonic function of x = log(concentration), we can convert confidence intervals for log(concentration) to ones for concentration by taking the same function of the endpoints of the interval for log(concentration). *Why is this okay? Can you explain it?* This yields an estimated concentration of $39.987 = e^{3.433}$ and a confidence interval of $[18.585, 51.666] = [e^{2.922}, e^{3.944}]$.

2. Cow pH example.

a) Showing SAS output.

```
                              Analysis of Variance
                                         Sum of           Mean
Source                        DF        Squares         Square    F Value    Pr > F
Model                          1        2.85695        2.85695     112.23    <.0001
Error                          8        0.20365        0.02546
Corrected Total                9        3.06060
             Root MSE                   0.15955    R-Square     0.9335
```

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.99649	0.09691	72.20	<.0001
time	1	-0.20869	0.01970	-10.59	<.0001

Covariance of Estimates

Variable	Intercept	time
Intercept	0.0093907462	-0.001629799
time	-0.001629799	0.0003880474

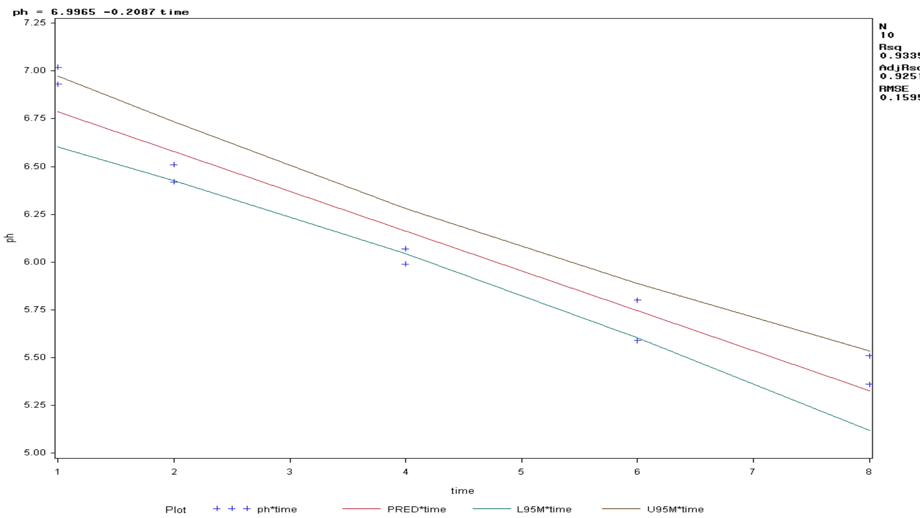

Figure 1: Cow pH; confidence intervals for the mean

b) This is is a regulation problem as we are trying to estimate that X, denoted $X(5.5)$, at which $E(Y|X) = m = 5.5$, which leads to $X(5.5) = (5.5 - \beta_0)/\beta_1$. From the output $b_0 = 6.99649$, $b_1 = -0.20869$, $s^2\{b_0\} = 0.0093907462$, $s\{b_0, b_1\} = -0.001629799$, $s^2\{b_1\} = 0.0003880474$ and $n = 10$. The point estimate is

$$\hat{X}(5.5) = (5.5 - b_0)/b_1 = 7.1709,$$

with an estimated standard error $s\{\hat{X}(m)\} = .3703 = \sqrt{.1371}$ computed using

$$s^2\{\hat{X}(m)\} = \frac{1}{b_1^2}[s^2\{b_0\} + \hat{X}(m)^2 s^2\{b_1\} + 2\hat{X}(m)s\{b_0, b_1\}$$

Using the estimated approximate standard error with $t(.975, 8) = 2.306$ yields the approximate interval
$\hat{X}(5.5) \pm t(.975, 8)SE = 7.1709 \pm 2.306 * .3703 = [6.3171, 8.0247]$.

Using Fieller's method the interval is computed with the formula in problem 1, but now $c_0 = (m - b_0)^2 - t^2 s^2\{b_0\}$, $c_{01} = (m - b_0)b_1 + t^2 s\{b_0, b_1\}$ and $c_1 = (b_1)^2 - t^2 s^2\{b_1\}$.

This yields and interval of [6.431, 8.206]. Note that because the p-value for the slope is less than .05 we know that the Fieller method will yield a finite interval.

Graphically, the Fieller interval is obtained by finding the points on the x-axis that correspond to where the CI for $E(\mu(x))$ equal 5.5 (or, said another way, drawing a line at y = 5.5, looking at where it intersects the CIs for the mean, E(Y), and projecting back to the x-axix).

Extra problem for S697R students.

At time X the pH Y is distributed $N(\beta_0 + \beta_1 X, \sigma^2)$ and $P(Y < 6) = P(Z < (6 - (\beta_0 + \beta_1 X))/\sigma)$ where Z is a standard normal random variable. If we set this equal to a specified probability π then we need $z_\pi = (6 - (\beta_0 + \beta_1 X))/\sigma)$, where z_π is the standard normal percentile with π to the left of it under the standard normal curve. Solving yields $\beta_0 + \beta_1 X = 6 - \sigma * z_\pi$ or $X = (6 - \sigma * z_\pi - \beta_0)/\beta_1$. This looks just like a regulation problem except m is replaced by $6 - \sigma * z_\pi$. If we know σ and have a specified π then we can proceed as in the regulation problem.

In the problem we want $P(Y > 6) = .05)$ so $\pi = .95$, $z_{.95} = 1.645$ and $6 - .01 * 1.645 = 5.98355$ (playing the role of m). This leads to
$\hat{X}(m) = 4.85380$
$SE = 0.2495$
and an approximate confidence interval of $[4.278, 5.429]$.

3. 2.39

   a) The marginal distribution of Y_1 is distributed normally with mean 50 and standard deviation 3 (or write as $Y_1 \sim N(50, 9)$).

   b) In general $Y_2|Y_1 = y_1$ is $N(\beta_0 + \beta_1 y_1, \sigma^2)$ (or in books notation $\beta_0 = \alpha_{21}$ and $\beta_1 = \beta_{21}$) where $\beta_1 = .8(4/3) = 1.0667$, $\beta_0 = 100 - 50(1.0667) = 46.67$ and $\sigma^2 = 5.76 = \sigma_2^2(1 - \rho_{12}^2) = 16(1 - .64) = \sigma_2^2 - \beta_1^2\sigma_1^2 = 16 - (1.0667^2)9 = 5.76$. So, $\sigma = 2.4$.

   So, $Y_2|Y_1 = y_1$ is $N(46.67 + 1.0667y_1, 5.76)$, i.e.

   Given $y_1 = 55$, Y_2 is $N(105.36, 5.76)$, or normal with mean 105 and standard deviation 2.4 (variance = 5.76).

   c) In general $Y_1|Y_2 = y_2$ is $N(\alpha_{12} + \beta_{12} y_2, \sigma^2)$ where $\beta_{12} = .8(3/4) = .6$, $\alpha_{12} = 50 - 100(.6) = -10$ and

   $\sigma^2 = 3.24 = \sigma_1^2(1 - \rho_{12}^2) = 9(1 - .64) = \sigma_2^2 - \gamma_1^2\sigma_1^2 9 - (.6 * *2)16 = 3.24 \qquad \sigma = 1.8.$

   $Y_1|Y_2 = y_2$ is $N(-10 + .6y_2, 3.24)$

   Given $y_2 = 95$, Y_1 is $N(47, 3.24)$ or normal with mean 47 and standard deviation 1.8.

4. - Problem 2.31. a) Here is the analysis of variance table along with the estimated coefficients, etc. (SAS output, R output very similar)

```
                              Analysis of Variance
                                      Sum of           Mean
Source                   DF          Squares         Square    F Value  Pr > F
Model                     1         93462942       93462942      16.83  <.0001
Error                    82        455273165        5552112
Corrected Total          83        548736108

                         Dependent Variable: rate
          Root MSE              2356.29195    R-Square     0.1703
          Dependent Mean        7111.20238    Adj R-Sq     0.1602

 Variable      DF       Estimate             Error     t Value     Pr > |t|
 Intercept      1          20518        3277.64269        6.26       <.0001
 hspercent      1     -170.57519          41.57433       -4.10       <.0001
```

b) **Not required** The F-test is given in the anova table. The F value of 16.83 is $= (-4.10)^2$, which is the t-statistics squared. Both tests lead to rejecting $H_0 : \beta_1 = 0$. The P-values are both given as $< .0001$. You could compute the exact p-values for each of these test as shown below. The differences are just due to rounding of statistics.

In SAS the probf and probt give cumulative probabilities for the F and t distributions respectively; i.e., $probf(c, d1, d2) = P(F \leq c)$ where F is distributed F with d1 and d2 degrees of freedom. In R the functions are pf and pt, respectively.

```
Using SAS

proc iml;
pvalue = 1 - probf(16.83,1,82);
pvaluet = 2*probt(-4.10,82);
print pvalue pvaluet;
run;
    pvalue   pvaluet
  0.0000959 0.0000967

Using R

pvalue = 1 - pf(16.83,1,82)
pvaluet = 2*pt(-4.1,82)
pvalue; pvaluet
[1] 9.587146e-05
[1] 9.671269e-05
```

c) $R^2 = .1703$ so 17.03 % of the total variation is explained or SSTO is reduced bye 17.03 %.

d) to compute r from the anova table we use $r = sign(b_1)\sqrt{.1703} = -\sqrt{.1703} = -.4127$.

Problems 2.48 and 2.49.

For 2.48, the sample Pearson correlation r is -0.41270. The p-value for testing the null hypothesis that the population correlation is zero is $< .0001$. We would reject H_0 and conclude that the population correlation is not 0. The p-value from this test (which is based on the assumption of bivariate normality) will be exactly the same as the P-value from the t-test (or F-test) for 0 slope in the regression of Y on X. You can't tell that they are exactly the same here where they are both listed as less than .0001, but see the Brain example for illustration.

For 2.49, the Spearman correlation is -.42593 with a P-value less than .0001. Very similar value and same conclusion as working with Pearson correlation.

NOTE: There is now a Fisher option in proc corr in SAS that does two things. Adjusts the estimate of correlation to account for potential bias and gives a 95% confidence interval for the population correlation ρ. This confidence interval depends on the bivariate normality assumption.

```
USING SAS

data a;
infile 'e:\crime.dat';
input rate hspercent;
proc reg;
model rate=hspercent;  run;
proc corr pearson spearman fisher;
var rate hspercent;  run;
proc univariate plot normal; var rate hspercent;
run;
                                The CORR Procedure
                    Pearson Correlation Coefficients, N = 84
                                          rate      hspercent
                     rate             1.00000       -0.41270
                                                      <.0001
                   Spearman Correlation Coefficients, N = 84
                                         rate      hspercent
                     rate             1.00000       -0.42593
                                                      <.0001
         Pearson Correlation Statistics (Fisher's z Transformation)
            With                   Sample                         Bias  Correlation
Variable   Variable          N  Correlation  Fisher's z  Adjustment     Estimate
rate       hspercent        84     -0.41270    -0.43887    -0.00249     -0.41064
                    With                                      p Value for
         Variable   Variable      95% Confidence Limits        H0:Rho=0
```

```
        rate       hspercent     -0.574459     -0.215188       <.0001
```

```
***USING R***

data<-read.table("f:/s505/data/crime.dat")
attach(data)
rate<-V1; hspercent<-V2
summary(rate)
summary(hspercent)
plot(hspercent,rate)
cor.test(rate,hspercent)   # will default to Pearson correlation
cor.test(rate,hspercent, method = c("spearman"))  # Does Spearman correlation

> summary(rate)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2105    5020    6930    7111    8840   14020
> summary(hspercent)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  61.00   76.00   79.00   78.60   82.25   91.00

> cor.test(rate,hspercent)   # will default to Pearson correlation

        Pearson's product-moment correlation
data:  rate and hspercent
t = -4.1029, df = 82, p-value = 9.571e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.5761223 -0.2175580
sample estimates:       cor  -0.4127033

> cor.test(rate,hspercent, method = c("spearman"))  # Does Spearman correlation

        Spearman's rank correlation rho
data:  rate and hspercent
S = 140839.3, p-value = 5.359e-05
alternative hypothesis: true rho is not equal to 0
sample estimates:       rho  -0.4259324
```