STAT 350

Assignment 1 Solutions

1. Suppose $Z_1, \ldots, Z_{10}$ are independent random variables each having a $N(0, 6)$ distribution. Let $\bar{Z} = \sum Z_i/10$, $U = \sum_{i=1}^4 Z_i^2/6$, $V = \sum_{i=5}^{10} Z_i^2/6$, $X = Z_1/\sqrt{V}$ and $Y = (3U)/(2V)$. Give the names for the distributions of each of $\bar{Z}$, U, V, X and Y and use tables to find $P(|\bar{Z}| > 1)$, $P(U \le 9.49)$, $P(-1.2 \le X \le 1.2)$, $P(Y > 6.23)$, $P(U \le V)$, $P(Z_1 - 2Z_2 \ge 10)$.

Solution

$\bar{Z}$ is $N(0, 6/10)$, U is χ_4^2, V is χ_6^2, X is t_6, since Z_1 and V are independent and $Z_1/\sqrt{V} = (Z_1/\sqrt{6})/\sqrt{V/6}$, and $Y = (U/4)/(V/6)$ has an $F_{4,6}$ distribution, using the fact that U and V are independent. Your answer should specify the mean and variance for $\bar{Z}$, the various degrees of freedom and note the required independence for X and Y. The required probabilities are 0.197, 0.95, 0.725, 0.025, 0.688, 0.034. I used SPlus to compute these; if you used tables your answers will be less accurate, particularly for $P(U \le V) = P(U/V \le 1) = P(Y < 1.5)$.

You need to remember:

(a) $Z_i/\sqrt{6}$ has a $N(0, 1)$ distribution.

(b) so $Z_i^2/6$ has a χ_1^2 distribution.

(c) $Var(Z_1 - 2Z_2) = 6 + (-2)^2 6 = 30$.

(d) $Var(Z_i) = 6$ (several people thought $Var(Z) = 6^2$).

2. A new process for measuring the concentration of a chemical in water is being investigated. A total of n samples are prepared in which the concentrations are the known numbers x_i for $i = 1, \ldots, n$; the new process is used to measure the concentrations for these samples. It is thought likely that the concentrations measured by the new process, which we denote Y_i, will be related to the true concentrations via

$$Y_i = \beta x_i + \epsilon_i$$

where the ϵ_i are independent, have mean 0 and all have the same variance σ^2 which is unknown.

(a) If this model is fitted by least squares, (that is by minimizing $\sum_i (Y_i - \beta x_i)^2$) show that the least squares estimate of β is

$$\hat{\beta} = \frac{\sum x_i Y_i}{\sum x_i^2}.$$

Solution

Differentiate $\sum_i (Y_i - \hat{\beta} x_i)^2$ with respect to $\hat{\beta}$ and get $-2\sum x_i(Y_i - \hat{\beta} x_i)$ which is 0 if and only if $\hat{\beta} = \sum x_i Y_i / \sum x_i^2$. The second derivative of the function being minimized is $2\sum x_i^2 > 0$ so this is a minimum.

Note: I saw a dismaying tendency to write things like

$$x_i \sum (Y_i - x_i \beta)$$

moving the x_i inside and out of the sum whenever you felt like it. Since it depends on i and the sum is over different values of i it is **not** a common factor in the sum so you can't take it outside the sum!

(b) Show that the estimator in part (a) is unbiased.

Solution

Let $K = 1/\sum x_i^2$. Then $E(\hat{\beta}) = KE(\sum x_i Y_i)$. Use $E(Y_i) = \beta x_i$ to see that

$$E(\hat{\beta}) = K \sum x_i^2 \beta = \beta.$$

(c) Compute (give a formula for) the standard error of $\hat{\beta}$.

Solution

You have to compute $Var(\hat{\beta}) = K^2 \sum Var(x_i Y_i) = \sigma^2 K$ which is simply $\sigma^2/\sum x_i^2$. The standard error is then $\sqrt{\mathrm{Var}(\hat{\beta})} = \sigma\sqrt{K}$.

(d) The error sum of squares for this model is $\sum (Y_i - \hat{\beta} x_i)^2$ which may be shown to have $n-1$ degrees of freedom. If the x_i are the numbers 1, 2, 3 and 4, $\hat{\beta} = 1$ and the error sum of squares is 0.12 find a 95% confidence interval for β and explain what further assumptions you must make to do so.

Solution

If we assume that the errors are independent $N(0, \sigma^2)$ random variables then $\hat{\beta}$ is independent of the usual estimate of σ^2, samely $s^2 = 0.12/3 = 0.04$ in this case. The usual t statistic then has a t distribution and the confidence interval is $1 \pm t_{3,0.025}\sqrt{0.04}/\sqrt{1^2 + 2^2 + 3^2 + 4^2}$ which boils down to $1 \pm (3.18)(0.2)/\sqrt{30}$.

Note: several people used formulas for simple linear regression as if they had fitted an intercept as well. They then had 2 degrees of freedom in spite of the fact that the question says there are $n-1$ degrees of freedom for error. Also: many of you neglected to assume that the errors have a normal distribution.

(e) Show that the estimator

$$\tilde{\beta} = \frac{\sum Y_i}{\sum x_i}$$

is also unbiased.

Solution

Let $K_2 = 1/\sum x_i$; then $E(\tilde\beta) = K_2 \sum x_i \beta = \beta$.

(f) Compute (give a formula for) the standard error of $\tilde\beta$. Which is bigger, the standard error of $\hat\beta$ or that of $\tilde\beta$?

Solution

We have $Var(\tilde\beta) = K_2^2 \sum Var(Y_i) = n\sigma^2/(\sum x_i)^2$. The difference $Var(\tilde\beta) - Var(\hat\beta)$ is then simply

$$\sigma^2 \frac{n\sum x_i^2 - (\sum x_i)^2}{\sum x_i^2 (\sum x_i)^2}$$

The denominator is positive and the numerator is $n(n-1)$ times the usual sample variance of the x's so this difference of variances is positive. This proves that the standard error of $\hat\beta$ is smaller than that of $\tilde\beta$. The standard error of $\tilde\beta$ itself is

$$\sqrt{\mathrm{Var}(\tilde\beta)} = \sigma \sqrt{\frac{n}{(\sum x_i)^2}}$$

(g) Show that the mle of β in this model is $\hat\beta$, the least squares estimate, if the ϵ_i have normal distributions.

Solution

In this case Y_i is $N(x_i\beta, \sigma^2)$ and the likelihood is

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-(Y_1 - x_1\beta)^2/(2\sigma^2)} \times \cdots \times \frac{1}{\sqrt{2\pi}\sigma} e^{-(Y_n - x_n\beta)^2/(2\sigma^2)}$$

As usual you maximize the logarithm which is

$$-n \log\sqrt{2\pi} - n\log\sigma - \frac{1}{2\sigma^2}\sum (Y_i - x_i\beta)^2$$

Take the β derivative and get the same equation to solve for β as in the first part of this problem.

3. Consider the two-way layout without replicates. We have data Y_{ij} for $i = 1, \ldots I$ and $j = 1, \ldots, J$. We generally fit a so-called additive model

$$Y_{ij} = \mu + \rho_i + \gamma_j + \epsilon_{ij}$$

In the following questions consider the case $I = 2$ and $J = 3$.

(a) If we treat μ, ρ_1, ρ_2, γ_1, γ_2 and γ_3 as the entries in the parameter vector β what is the design matrix $X = X_a$ and what is the rank of X_a?

Writing the data as $Y^T = [Y_{1,1}, Y_{1,2}, Y_{1,3}, Y_{2,1}, Y_{2,2}, Y_{2,3}]$ the design matrix is

$$X_a = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Letting x_i denote column i of X_a we have $x_3 = x_1 - x_2$ and $x_6 = x_1 - x_4 - x_5$ so that the rank of X_a must be no more than 4. But if $a_1x_1 + a_2x_2 + a_4x_4 + a_5x_5 = 0$ then from row 6 we get $a_1 = 0$. Then from rows 4 and 5 get $a_4 = 0$ and $a_5 = 0$. Finally use row 3 to get $a_2 = 0$. This shows that columns 1, 2 4 and 5 are linearly independent so tha tX_a has rank at least 4 and so exactly 4.

(b) What is the determinant of the matrix $X_a^T X_a$? Is this matrix invertible? How many solutions do the normal equations have?

The matrix $X_a^T X_a$ is 6 by 6 but has rank only 4 so is singular and must have determinant 0. The normal equations are

$$\begin{bmatrix} 6 & 3 & 3 & 2 & 2 & 2 \\ 3 & 3 & 0 & 1 & 1 & 1 \\ 3 & 0 & 3 & 1 & 1 & 1 \\ 2 & 1 & 1 & 2 & 0 & 0 \\ 2 & 1 & 1 & 0 & 2 & 0 \\ 2 & 1 & 1 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} \mu \\ \rho_1 \\ \rho_2 \\ \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{bmatrix} = \begin{bmatrix} \sum_{ij} Y_{ij} \\ \sum_j Y_{1j} \\ \sum_j Y_{2j} \\ \sum_i Y_{i1} \\ \sum_i Y_{i2} \\ \sum_i Y_{i3} \end{bmatrix}$$

It may be seen that the second and third rows give equations which add up to the equation in the first row as do the fourth, fifth and sixth rows. Eliminate rows 3 and 6, say and solve. This leaves 4 linearly independent equations in 6 unknowns and so there are infinitely many solutions.

NOTE: You need to check that the equations are consistent. Many students row reduced only the matrix of coefficients forgetting the fact that if A has less than full rank $Ax = b$ might have 0 solutions. My solution above row reduces the **augmented** matrix, checking that there is at least one solution..

(c) Usually we impose the restrictions $\rho_1 + \rho_2 = 0$ and $\gamma_1 + \gamma_2 + \gamma_3 = 0$. Use these restrictions to eliminate ρ_2 and γ_3 from the model equation and, for the parameter vector $\beta^T = (\mu, \rho_1, \gamma_1, \gamma_2)$ find the design matrix X_b.

The restrictions give $\rho_2 = -\rho_1$ and $\gamma_3 = -\gamma_1 - \gamma_2$. In each model equation which mentions either ρ_2 or γ_3 you replace that parameter by the equivalent formula. So, for example,

$$\begin{aligned} Y_{2,3} &= \mu + \rho_2 + \gamma_3 + \epsilon_{2,3} \\ &= \mu - \rho_1 - \gamma_1 - \gamma_2 + \epsilon_{2,3} \end{aligned}$$

The 6th row of the design matrix is obtained by reading off the coefficients of $\mu, \rho_1, \gamma_1, \gamma_2$ which are 1, -1, -1 and -1. This makes

$$X_b = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix}$$

Note: some people eliminated ρ_1 and not ρ_2. This is fine and leads to a similar matrix with the rows in a different order.

(d) An alternate set of restrictions is called corner point coding where we assume $\rho_1 = \gamma_1 = 0$. With this restriction and the parameter vector $\beta^T = (\mu, \rho_2, \gamma_2, \gamma_3)$ what is the design matrix X_c?

This just makes the design matrix X_c just the corresponding columns, 1, 3, 5 and 6 of X_a.

(e) Show that the three design matrices have the same column space by finding a matrix A such that $X_a = X_b A$ and similarly for X_b and X_c and for X_a and X_c.

To write $X_c = X_a A$ just let A be the 6×4 matrix which picks out columns 1,3,5 and 6 of X_a, namely,

$$A_{ca} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

To write $X_a = X_c A$ we just have to put back column 2 and 4 remembering that col 2 is col 1 - col 3 and col 4 is col 1 - col 5 - col 6. Thus

$$A_{ac} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

Similarly

$$A_{bc} = \begin{bmatrix} 1 & 1/2 & 1/3 & 1/3 \\ 0 & -1/2 & 0 & 0 \\ 0 & 0 & -1/3 & -1/3 \\ 0 & 0 & 2/3 & -1/3 \end{bmatrix}$$

A vector in the column space of say X_a is of the form $X_a v$ for a vector of coefficients v. But such a vector is $X_c A_{ac} v$ and so of the form $X_c w$ for the vector $w = A_{ac} v$ and so in the column space of X_c.

(f) Use the previous part to show that the vectors of fitted values $\hat{Y}$ will be the same for any solution of the normal equations for any of the three design matrices.

The easy way to do this is to say: the fitted vector $\hat{\mu}$ is the closest point in the column space of the design matrix to the data vector Y. Since all three have the same column space they all have the same closest point and so the same $\hat{\mu}$.

Algebra is an alternative tactic: The matrix A_{bc} is invertible and we have

$$\hat{\mu}_b = X_b\hat{\beta}_b = X_b(X_b^TX_b)^{-1}X_b^TY$$

Plug in X_cA_{bc} for X_b and get

$$\hat{\mu}_b = X_cA_{bc}(A_{bc}^TX_c^TX_cA_{bc})^{-1}A_{bc}^TX_c^TY$$

Use $(BC)^{-1} = C^{-1}B^{-1}$ to see that all occurences of A_{bc} cancel out to give

$$\hat{\mu}_b = X_c(X_c^TX_c)^{-1}X_c^TY = \hat{\mu}_c\,.$$

The algebraic approach makes it a bit more difficult to deal with the case of X_a because the normal equations have many solutions. Suppose that $\tilde{\beta}$ is any solutions of the normal equations $X_a^TX_a\tilde{\beta} = X_a^TY$. Then

$$A_{ac}^TX_c^TX_cA_{ac}\tilde{\beta} = A_{ac}^TX_c^TY$$

or

$$A_{ac}^T\left(X_c^TX_cA_{ac}\tilde{\beta} - X_c^TY\right) = 0$$

The matrix A_{ac} has rank 4. If any vector v satisfies $A_{ac}^Tv = 0$ then

$$A_{ac}^Tv = \begin{bmatrix} v_1 \\ v_1 - v_2 \\ v_2 \\ v_1 - v_3 - v_4 \\ v_3 \\ v_4 \end{bmatrix}$$

so that $v_1 = v_2 = v_3 = v_4 = 0$. We apply this with

$$v = X_c^TX_cA_{ac}\tilde{\beta} - X_c^TY\,.$$

This shows that

$$X_c^TX_cA_{ac}\tilde{\beta} = X_c^TY$$

so that $A_{ac}\tilde{\beta} = \hat{\beta}_c$.

Thus

$$\hat{\mu}_a = X_a\tilde{\beta} = X_cA_{ac}\tilde{\beta} = X_c\hat{\beta}_c = \hat{\mu}_c$$

4. From the text question 1.19, 1.23, 2.13 a and b and 2.23 a, b and c. In 2.23 c give a P-value and interpret this P-value.

The following SAS code was used for all these parts:

```
data gpa;
 infile "CH01PR19.DAT";
 input gpa test;
proc glm;
  model gpa=test;
  estimate "fit_at_5" intercept 1 test 5;
  estimate "fit4.7" intercept 1 test 4.7;
  output out=gpaout r=resid p=fitted ;
run;
proc print data=gpaout;
  var test gpa fitted resid;
run;
proc means data=gpaout;
 var resid;
run;
proc rank normal=vw data=gpaout out=gpaout2;
  var resid;
  ranks normscr;
run;
proc gplot data=gpaout2;
 plot gpa*test fitted *test /overlay;
run;
proc gplot data=gpaout2;
 plot fitted*test;
 plot resid*normscr;
run;
```

obtaining the following output:

```
Dependent Variable: GPA
                                     Sum of            Mean
Source                  DF          Squares          Square  F Value    Pr > F
Model                    1       6.43372807      6.43372807    34.00    0.0001
Error                   18       3.40627193      0.18923733
Corrected Total         19       9.84000000

                  R-Square             C.V.        Root MSE           GPA Mean
                  0.653834         17.40057         0.43501            2.50000

Source                  DF         Type I SS     Mean Square  F Value    Pr > F
TEST                     1       6.43372807      6.43372807    34.00    0.0001
Source                  DF       Type III SS     Mean Square  F Value    Pr > F
TEST                     1       6.43372807      6.43372807    34.00    0.0001
```

```
Dependent Variable: GPA
                                   T for H0:     Pr > |T|    Std Error of
Parameter           Estimate     Parameter=0                   Estimate
fit_at_5          2.50000000           25.70       0.0001      0.09727213
fit4.7            2.24802632           21.12       0.0001      0.10643937

                                   T for H0:     Pr > |T|    Std Error of
Parameter           Estimate     Parameter=0                   Estimate
INTERCEPT       -1.699561404           -2.34       0.0311      0.72677682
TEST             0.839912281            5.83       0.0001      0.14404759

                 OBS    TEST    GPA     FITTED       RESID

                   1     5.5    3.1    2.91996     0.18004
                   2     4.8    2.3    2.33202    -0.03202
                   3     4.7    3.0    2.24803     0.75197
                   4     3.9    1.9    1.57610     0.32390
                   5     4.5    2.5    2.08004     0.41996
                   6     6.2    3.7    3.50789     0.19211
                   7     6.0    3.4    3.33991     0.06009
                   8     5.2    2.6    2.66798    -0.06798
                   9     4.7    2.8    2.24803     0.55197
                  10     4.3    1.6    1.91206    -0.31206
                  11     4.9    2.0    2.41601    -0.41601
                  12     5.4    2.9    2.83596     0.06404
                  13     5.0    2.3    2.50000    -0.20000
                  14     6.3    3.2    3.59189    -0.39189
                  15     4.6    1.8    2.16404    -0.36404
                  16     4.3    1.4    1.91206    -0.51206
                  17     5.0    2.0    2.50000    -0.50000
                  18     5.9    3.8    3.25592     0.54408
                  19     4.1    2.2    1.74408     0.45592
                  20     4.7    1.5    2.24803    -0.74803

        Analysis Variable : RESID

         N          Mean       Std Dev       Minimum       Maximum
        ----------------------------------------------------------
        20  -7.77156E-17     0.4234117    -0.7480263     0.7519737
        ----------------------------------------------------------
```

Now to answer the questions:

1.19 a) The estimates are $\hat{\beta}_0 = -1.7$ and $\hat{\beta}_1 = 0.84$, rounded off sensibly.

b) It is not so easy to plot the data and regression line; the code above does so but the fitted line is plotted only at the available x values. However, nothing in the plot suggests any big problems.

c) The first `estimate` line produces the estimae 2.5 with a standard error of 0.097.

d) This just asks for the slope again which is 0.84.

1.23 a) The `proc print` prints out the residuals and the `proc means` shows that the variable `resid` has mean very close to 0.

b) The MSE is $\hat{\sigma}^2 = 0.189$ and the estimate of σ is the Root MSE, 0.435. The RMSE has the same units as σ which has the same units as Y, namely, grade points.

2.13 a) This confidence interval is

$$\hat{\beta}_0 + 4.7\hat{\beta}_1 \pm t_{0.025,18}\hat{\sigma}\sqrt{x^T(X^TX)^{-1}x}$$

where $x^T = (1, 4.7)$. The quantity $\hat{\sigma}\sqrt{x^T(X^TX)^{-1}x}$ is printed out by the `estimate` statement. It is the standard error 0.106. The actual estimate is 2.248 and the t multiplier is 2.101 so the interval is

$$2.248 \pm 2.101 \times 0.106 \qquad \text{or} \qquad (2.025, 2.471)$$

with one more digit than needed just so you can check your method.

b) The prediction for an individual is the same but the interval is

$$\hat{\beta}_0 + 4.7\hat{\beta}_1 \pm t_{0.025,18}\hat{\sigma}\sqrt{1 + x^T(X^TX)^{-1}x}$$

or

$$2.248 \pm 2.101 \times \sqrt{0.106^2 + 0.189}$$

or just

$$(1.308, 3.188)$$

Notice that this is VERY wide.

2.23 a) The ANOVA table is straight from the output:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6.43372807	6.43372807	34.00	0.0001
Error	18	3.40627193	0.18923733		
Corrected Total	19	9.84000000			

b) MSR is an estimate of its expected value. See formula 2.57 in the text:

$$\mathrm{E}(MSE) = \sigma^2 + \beta_1^2 \sum (x_i - \bar{X})^2$$

You can use, for instance, `proc means` to compute the sum of squares at the end or algebraic identities to deduce it from `proc glm` output. The MSE is an unbiased estimate of σ^2. These two expected values are equal if the hypothesis $\beta_1 = 0$ is true.

c) The F statistic in question is printed in the ANOVA table. It is 34 and the P-value is 0.0001 so that at any reasonable level the hypothesis that $\beta_1 = 0$ is rejected. The null hypothesis is $\beta_1 = 0$, the alternative is $\beta_1 \neq 0$ and the conclusion is that the true slope is NOT 0.