

Stat312: Final Exam Solutions

Moo K. Chung
mchung@stat.wisc.edu

December 17, 2004

1. We wish to fit data $(x_1, y_1), \dots, (x_n, y_n)$ with linear model $Y_j = 1 + \beta x_j + \epsilon_j$ with $\epsilon_j \sim N(0, \sigma^2)$. Derive everything.
- (a) Find the maximum likelihood estimator for β (10pts).
 - (b) Estimate β by the least squares method (10pts).
 - (c) Determine the distribution of the estimator given in (b) (10pts).

Solution.

(a) The density $f(\epsilon_j) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\epsilon_j^2/2\sigma^2}$. The density $f(y_j) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_j-1-\beta x_j)^2/2\sigma^2}$. The likelihood function is

$$L(\beta) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\sum_{j=1}^n (y_j-1-\beta x_j)^2/2\sigma^2}$$

and the loglikelihood is

$$\ln L(\beta) = \text{const.} - n \ln \sigma - \sum_{j=1}^n (y_j - 1 - \beta x_j)^2 / 2\sigma^2.$$

$$\partial \ln L(\beta) / \partial \beta = \sum_{j=1}^n x_j (y_j - 1 - \beta x_j) / \sigma^2 = 0.$$

Solving this we get the same result as (b).

(b) Let $z_j = y_j - 1$. Then the problem is equivalent to a regression $Z_j = \beta x_j + \epsilon_j$. The sum of squared errors $SSE = \sum_{i=1}^n (z_j - \hat{\beta} x_i)^2$ where $\hat{\beta}$ is an estimator of β . Letting $\partial SSE / \partial \hat{\beta} = 0$, we get $-2 \sum_{i=1}^n x_i (z_i - \hat{\beta} x_i) = 0$. Solving this,

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i z_i}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2} = \frac{\overline{xy} - \bar{x}}{\overline{x^2}}.$$

(c) Since $Y_i \sim N(1 + \beta x_j, \sigma^2)$, our estimator should be also normally distributed. Note that

$$\mathbb{E}\hat{\beta} = \frac{\sum_{i=1}^n x_i \mathbb{E}Y_i - \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2} = \frac{\sum_{i=1}^n x_i (1 + \beta x_j) - \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2} = \beta,$$

$$\mathbb{V}\hat{\beta} = \mathbb{V}Y_i = \sigma^2.$$

Hence $\hat{\beta} \sim N(\beta, \sigma^2)$.

2. Let $X_1, \dots, X_m \sim N(\mu_1, \sigma^2)$ and $Y_1, \dots, Y_m \sim N(\mu_2, \sigma^2)$. Further two samples are independent. Estimate μ_1, μ_2 and σ^2 via the maximum likelihood method (15pts).

Solution. This problem was your 1st homework problem.

3. A manufacturer of automatic washers offers a particular model in one of two colors: white and green. Of the 10 washers sold, 7 were white and 3 were green. Would you conclude that customers have a preference for white color? Test it at $\alpha = 0.05$. Also compute the p -value corresponding to the test procedure (15 pts).

Solution. Let p and q be the proportion of customers who prefer white or green respectively. Obviously $p + q = 1$. We test

$$H_0 : p = 0.5 \text{ vs. } H_1 : p > 0.5$$

Since the sample size is too small, our test statistic will be based on Binomial distribution. Let X_i be a Bernoulli random variable defined as $X_i = 1$ if the i -washer sold is white. Then $P(X_i = 1) = p$. The test statistic is $X = \sum_{i=1}^{10} X_i \sim \text{Bin}(10, 0.5)$ under H_0 . We reject H_0 if X is large. Note that

$$P(X > 9) = P(X \geq 10) = 0.5^{10} = 0.0009765625,$$

$$P(X > 8) = P(X \geq 9) = \left[\binom{10}{9} + 1 \right] 0.5^{10} = 0.01074219,$$

$$P(X > 7) = P(X \geq 8) = \left[\binom{10}{8} + \binom{10}{9} + 1 \right] 0.5^{10} = 0.0546875.$$

Since we reject H_0 if $X > 8$ at $\alpha = 0.05$ So we do not reject H_0 . p -value is 0.0107.

4. Suppose that 20 out of 50 females and 10 out of 40 males are observed to be depressed in a psychological test. Does this data indicate the proportion of depression is different for the two groups? Test it at $\alpha = 0.05$ State your hypothesis and a test statistic and its distribution under the null hypothesis (10pts).

Solution. Let p_1 and p_2 be the proportions of depression in females and males respectively. Let X and Y be the number of depressed people in each group and $m = 50$ and $n = 40$. Then we test

$$H_0 : p_1 = p_2 \text{ vs. } H_1 : p_1 \neq p_2.$$

The test statistic is

$$Z = \frac{X/m - Y/n - (p_1 - p_2)}{\sqrt{p(1-p)(1/m + 1/n)}} = \frac{X/m - Y/n}{\sqrt{p(1-p)(1/m + 1/n)}} \sim N(0, 1)$$

under the null hypothesis. Since p is unknown, you estimate it by pooling the samples: $\hat{p} = (x + y)/(m + n) = (20 + 10)/(50 + 40) = 1/3$. $z = (0.4 - 0.25)/\sqrt{\frac{1}{3} \frac{2}{3} (1/50 + 1/40)} = 1.5$ We reject H_0 if $|z| > z_{0.5} = \text{qnorm}(0.95) = 1.64$. We reject H_0 and conclude that the proportion of depression is different between the groups.

5. 10 students took two midterm exams and the scores are listed here.

Midterm 1 || 80 75 60 90 99 60 55 85 65 70

Midterm 2 || 70 60 70 72 95 66 60 80 70 60

You may use the following R-output for answering the questions.

```
> x<-c(80, 75, 60, 90, 99, 60, 55, 85, 65, 70)
> y<-c(70, 60, 70, 72, 95, 66, 60, 80, 70, 60)
> summary(lm(y~x))
```

```
Call: lm(formula = y ~ x)
```

```
Residuals:
```

Min	1Q	Median	3Q	Max
-10.908	-6.312	1.758	4.354	10.836

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	29.4827	13.2261	2.229	0.0564 .
x	0.5523	0.1760	3.139	0.0138 *

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 7.647 on 8 degrees of freedom Multiple  
R-Squared: 0.5519, Adjusted R-squared: 0.4959 F-statistic: 9.854  
on 1 and 8 DF, p-value: 0.01383
```

- (a) Test if the two midterm scores are statistically independent at $\alpha = 0.05$. What statistical assumptions are you making to perform the test? (5pts)
- (b) Test if the second midterm exam is easier than the first midterm exam at $\alpha = 0.05$. What statistical assumptions are you making to perform the test? (10pts)

Solution. (a) Testing independence of two variables is equivalent to testing on the coefficient of the regression under the bivariate normal assumption. So we test

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0.$$

The test is based on t statistic with 8 degrees of freedom and the resulting t value is 3.139 and the corresponding p value is 0.0138. So we reject H_0 and conclude that the two test scores are not independent.

(b) The test will be based on t statistic for a paired data. To perform t test, we assume the data is coming from a normal distribution with unknown variance. Let μ_1 be the mean score for the first midterm and μ_2 be the mean score for the second midterm. Then we are interested in testing

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_1 : \mu_1 > \mu_2.$$

let $w_i = x_i - y_i$ and $\mu = \mu_1 - \mu_2$. Then we are testing

$$H_0 : \mu = 0 \text{ vs. } H_1 : \mu > 0.$$

Then using calculator, we can compute the sample mean and the sample standard deviation to be $\bar{w} = 3.6, S_w = 9.698$. Then the test stat is $t = \bar{w}/(S_w/\sqrt{10}) = 1.1739$. The critical threshold is $t_{0.05,9} = \text{qt}(0.95, 9) = 1.833$. We do not reject H_0 so we conclude that the second midterm is not harder than the first one at $\alpha = 0.05$.