

Math 218 Mathematical Statistics

Prof. D. Joyce, Clark University

20 Mar 2008

Second Test. Wednesday, 25 Mar 2009. On chapters 6–9.

Due Today. From chapter 8, exercise 14, and from Chapter 9, exercises 1–3, 6.

After the test. Presentations by you. Today we'll divide the class up into groups and each group will work on one of the problems 4, 5, 6, or 7 from chapter 10, pages 387–389. There's a fair amount of work to do for each problem ranging from computation to presentation, so decide who's going to do what and how you're going to present it. The computations could be done by a hand calculator, but are probably better done with a spreadsheet or, even better, with a statistics package. You could give the presentation with printed handouts, spreadsheets projections, or whatever you like.

Last time. Analysis of the model for simple linear regression

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

including SST (sum of the squared errors), SST (total sum of squares), and SSR (regression sum of squares), where

$$\text{SST} = \text{SSR} + \text{SSE}$$

and a bit on correlation.

$$r^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}.$$

Furthermore $r = \hat{\beta}_1 s_x / s_y$.

Today. Statistical inferences based on the model.

We have three estimators $\hat{\beta}_0$, $\hat{\beta}_1$, and $S^2 =$

$\text{SSE}/(n-1)$ for the three unknown parameters β_0 , β_1 , and σ^2 . The first two are normal distributions with means being the parameters they're estimating and standard deviations

$$\text{SD}(\hat{\beta}_0) = \sigma \sqrt{\frac{\sum x_i^2}{n S_{xx}}}$$
$$\text{SD}(\hat{\beta}_1) = \frac{\sigma}{\sqrt{S_{xx}}}$$

so we can use them to make inferences about β_0 and β_1 . If σ happens to be known, or if n is large, we can standardize them and make z -tests and z -confidence intervals.

But if n is small, we'll need t -tests. In order to do that, we'll have to replace the unknown standard deviation σ by the sample standard deviation s , so the standard deviations $\text{SD}(\hat{\beta}_0)$ and $\text{SD}(\hat{\beta}_1)$ are replaced by estimated standard deviations

$$\text{SE}(\hat{\beta}_0) = s \sqrt{\frac{\sum x_i^2}{n S_{xx}}}$$
$$\text{SE}(\hat{\beta}_1) = \frac{s}{\sqrt{S_{xx}}}$$

to get t -distributions with $(n-2)$ degrees of freedom. Precisely,

$$\frac{\hat{\beta}_0 - \beta_0}{\text{SE}(\hat{\beta}_0)} \sim t_{n-2} \quad \text{and} \quad \frac{\hat{\beta}_1 - \beta_1}{\text{SE}(\hat{\beta}_1)} \sim t_{n-2}.$$

For example, the two-sided confidence intervals for β_0 and β_1 have endpoints

$$\hat{\beta}_0 \pm t_{n-2, \alpha/2} \text{SE}(\hat{\beta}_0) \quad \text{and} \quad \hat{\beta}_1 \pm t_{n-2, \alpha/2} \text{SE}(\hat{\beta}_1).$$

Estimating the error variance σ^2 of the model. As you would expect, some sort of sample variance ought to be an estimator for σ^2 . What works is

$$s^2 = \frac{\sum e_i^2}{n-2} = \frac{\text{SSE}}{n-2}$$

which is an unbiased estimator of σ^2 . It is an unbiased estimator for σ^2 . Scaling it to

$$\frac{(n-2)S^2}{\sigma^2} = \frac{\text{SSE}}{\sigma^2}$$

gives us a statistic with a χ^2 distribution with $n-2$ degrees of freedom. (Note how this shows dividing by $n-2$ or n doesn't affect computations since whichever is used, it has to be scaled away to get the χ^2 distribution.)

Degrees of freedom. When we have n data values y_1, \dots, y_n , we've got a point $\mathbf{y} = (y_1, \dots, y_n)$ in \mathbf{R}^n , and that point can be any point. There are n degrees of freedom in specifying \mathbf{y} .

If we translate these all by the sample mean \bar{y} to the values $y_1 - \bar{y}, \dots, y_n - \bar{y}$, we also get a point $\mathbf{u} = (u_1, \dots, u_n) = (y_1 - \bar{y}, \dots, y_n - \bar{y})$ in \mathbf{R}^n , but it can't be just any point in \mathbf{R}^n because its coordinates satisfy the equation $\sum u_i = 0$. In other words, the point \mathbf{u} lies in a hyperplane of \mathbf{R}^n , that is, a linear subspace of dimension $n-1$. So there are $n-1$ degrees of freedom in specifying \mathbf{u} .

This means that the total sum of squares $\text{SST} = \sum (y_i - \bar{y})^2$ has $n-1$ degrees of freedom since it is a function of a point (the \mathbf{u} above) that has $n-1$ degrees of freedom.

The error sum of squares $\text{SSE} = \sum \epsilon_i^2 = \sum (y_i - \hat{y}_i)^2$ turns out to have $n-2$ degrees of freedom since the point $\mathbf{v} = (y_1 - \hat{y}_1, \dots, y_n - \hat{y}_n)$ satisfies two equations in its coordinates. The regression sum of squares SSR has only 1 degree of freedom.

Mean square error (MSE) and Mean square regression (MSR).

At the moment the MSE and MSR aren't so important, but there's a connection to the t -statistic mentioned above. They'll get more interesting when we do multiple regression.

These are just the SSE and SSR divided by their degrees of freedom. We're doing simple regression

right now, so SSR has only 1 degree of freedom. But later we'll do multiple regression where there are k independent variables, and there SSR will have k degrees of freedom instead of just 1 degree of freedom, while SSE will have $n - (k+1)$ degrees of freedom instead of $n-1$ degrees of freedom.

The ratio $\frac{\text{MSR}}{\text{MSE}}$ is a square of that t -statistic mentioned above.

$$\begin{aligned} \frac{\text{MSR}}{\text{MSE}} &= \frac{\text{SSR}}{s^2} = \frac{\hat{\beta}_1^2 S_{x,x}}{s^2} \\ &= \left(\frac{\hat{\beta}_1}{s/\sqrt{S_{x,x}}} \right)^2 \\ &= \left(\frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} \right)^2 = t^2 \end{aligned}$$

Furthermore, the square of a t -statistic is an F -statistic, specifically, an $F_{1,\nu}$ -statistic. When we look at multiple regression, we'll see some of these 1's will be replaced by k 's.