CLARK UNIVERSITY

Inferences for simple linear regression
Math 218, Mathematical Statistics

D Joyce, Spring 2016

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

Next, we'll estimate the error variance σ^2 of the model. We'll start looking at statistical inferences based on the simple linear regression model.

Estimating σ^2. There are three parameters in the simple linear regression model, β_1, β_2, and σ^2. We've already got estimators for the first two. We need an estimator for σ^2.

As you would expect, some sort of sample variance ought to do it. Such a thing is

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

This is an unbiased estimator of σ^2.

More importantly, for statistical inferences, by appropriately scaling the sample variance, we get a χ^2 distribution

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

(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.)

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

$$\text{SD}(\beta_0) = \sigma\sqrt{\frac{\sum x_i^2}{nS_{xx}}}$$

$$\text{SD}(\beta_1) = \frac{\sigma}{\sqrt{S_{xx}}}$$

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 $\text{SD}(\beta_0)$ and $\text{SD}(\beta_1)$ are replaced by estimated standard deviations

$$\text{SE}(\beta_0) = s\sqrt{\frac{\sum x_i^2}{nS_{xx}}}$$

$$\text{SE}(\beta_1) = \frac{s}{\sqrt{S_{xx}}}$$

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

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

The third estimator S^2 can be scaled to have a χ^2 distribution with $n-2$ degrees of freedom

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

Degrees of freedom. When we have n data values $y_1, \ldots, y_n$, we've got a point $\mathbf{y} = (y_1, \ldots, y_n)$ 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 $\overline{y}$ to the values $y_1 - \overline{y}, \ldots, y_n - \overline{y}$, we also get a point $\mathbf{u} = (u_1, \ldots, u_n) = (y_1 - \overline{y}, \ldots, y_n - \overline{y})$ 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 $\text{SST} = \sum(y_i - \overline{y})$ 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 $\text{SSE} = \sum \epsilon_i = \sum(y_i - \hat{y}_i)^2$ turns out to have $n-2$ degrees of freedom since the point $\mathbf{v} = (y_1 - \hat{y}_1, \ldots, y_n - \hat{y}_n)$ 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.

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

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.

Confidence and prediction intervals for simple linear regression. In the model for simple linear regression

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

different values of x produce different predictions for $Y = \beta_0 + \beta_1 x + \epsilon$. After getting n data values, we compute the least squares line

$$y = \hat{\beta}_0 + \hat{\beta}_1 x,$$

and we can determine confidence intervals for the parameters β_0, β_1, and σ^2.

But, we can do more. Suppose we set the predictor variable x to some specified value, x^*. (It would also be reasonable to denote the new values with a subscript of $n+1$ rather than a superscript of $*$.) That introduces a new Y value

$$Y^* = \beta_0 + \beta_1 x^* + \epsilon^*$$

where ϵ^* is a new independent error random variable with the same normal distribution as the other ϵ_i's, namely, $\text{NORMAL}(0, \sigma^2)$. Therefore, the random variable Y^* is $\text{NORMAL}(\beta_0 + \beta_1 x^*, \sigma^2)$.

Our model gives a predicted value of Y^*:

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*.$$

We'll denote the mean of Y^* by μ^*, thus

$$\mu^* = E(Y^*) = \beta_0 + \beta_1 x^*.$$

We don't know what μ^* is since we don't know β_0 and β_1, but we have a predicted value for it, which is the same as the predicted value for Y^*:

$$\hat{\mu}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*.$$

Although these predicted values are the same, when we use them to determine intervals for what they predict, namely, y^* and μ^*, respectively, we'll get different width intervals since the variances are much smaller for the mean.

A *prediction interval* (PI) for Y^* at α is an interval centered at $\hat{y}^*$ such that the probability that Y^* lies in that interval is $100(1-\alpha)\%$. It works out to be that its endpoints are

$$\hat{y}^* \pm t_{n-2,\alpha/2}\, s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}$$

where $s = \sqrt{\text{MSE}}$ is the estimator of σ as we've saw before.

This interval is closely related to the confidence interval for $\mu^* = \beta_0 + \beta_1 x^*$. At the significance level α, it has endpoints

$$\hat{\mu}^* \pm t_{n-2,\alpha/2}\, s\sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}.$$

If you graph the endpoints of either of these two intervals with x^* on the x-axis, you'll get a pair of hyperbolas, one above the least squares line, the other equally far below it. When x^* is close to $\overline{x}$, the hyperbolas are close, and that indicates that the prediction interval for Y^* and the confidence interval for μ^* are shorter there. But when x^* is far from $\overline{x}$, the hyperbolas spread apart, so the intervals there are large. In other words, good predictions can be made near $\overline{x}$, not so good predictions far from $\overline{x}$.

The prediction interval for the next observation Y^* is, of course, much larger than the confidence interval for for the mean μ^*. Graphically, the prediction hyperbola for Y^* are much further away from the least squares line than the hyperbola for μ^*.

Regression diagnostics. A basic question is: do the data support the hypotheses necessary for the simple linear regression model? A preliminary test we've already seen is to make a scatter plot of the data $(x_1, y_1), \ldots, (x_n, y_n)$. If it's obvious that the plot is nonlinear, then maybe the model is not appropriate.

Other tests can be made after fitting the least squares line, and some of these depend on making a scatter plot of the residuals $e_i = y_i - \hat{y}_i$, that is the plot of $(x_1, e_1), \ldots, (x_n, e_n)$.

If the hypotheses for the regression model is correct, then the e_i's are normally distributed with mean 0 and variance close to (but a little less than) σ^2. They are not independent since $\sum e_i = 0$ and $\sum x_i e_i = 0$ as shown in the text.

If the residual scatter plot shows some pattern, then the simple linear model may not be the best. That's not necessarily a bad thing as the pattern may indicate better models.

For instance, in the tread wear example in the text, there's a clear parabolic form to the residual plot. That suggests a quadratic model might be better. We'll see after we've introduced multiple linear regression that quadratic models can be subsumed in those models, so we'll see this example later.

There are other things that the residual plot can show, but it may take a large n to see them. The model assumes that the errors ϵ_i have the same variance for all i. If in the plot it appears that the residuals are close to 0 at one end, but scatter far from 0 at the other, then a transformation may be needed before applying the model, that is, some linearizing transformation needs to be applied to the y-values to make the variance of the resulting errors more uniform across all the x-values. Examples in the text use the log function and the reciprocal function as transformations.

Math 218 Home Page at
`http://math.clarku.edu/~djoyce/ma218/`