Simple Linear Regression and Correlation (Devore Chapter Twelve)

MATH-252-01: Probability and Statistics II*

Spring 2019

Contents

Tuesday 26 March 2019

1 Linear Regression

We make a bit of a shift of topic now, from statistical tests and confidence intervals to the process of fitting data to a model. The techniques will seem rather different, but it's enlightening to make contact with what we've done before.

Recall that in our treatment of paired data, we considered data sets $\{x_i | i = 1, \ldots, n\}$ and $\{y_i | i = 1, \ldots, n\}$ which were assumed to drawn from joint distributions where X_i and the corresponding Y_i were correlated (for instance a bivariate normal distribution), and we drew inferences on the statistical properties of

$$D_i = X_i - Y_i \tag{1.1}$$

We can write this relationship as

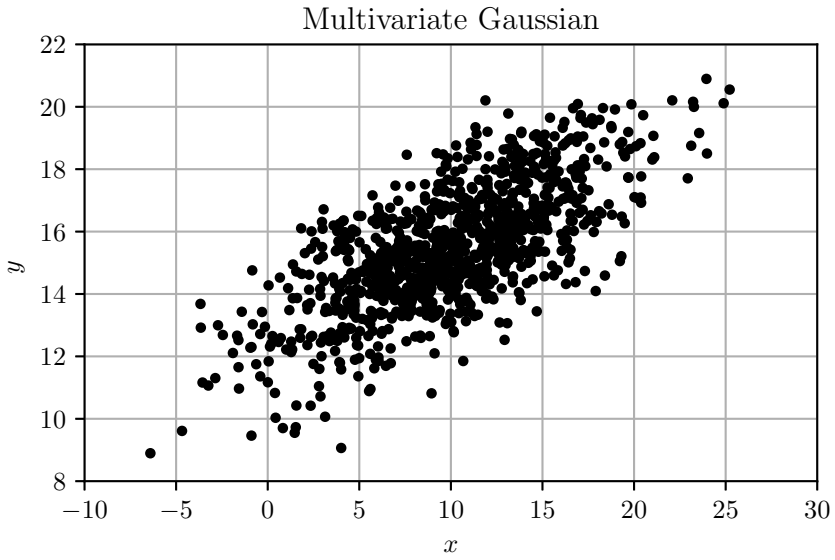
$$Y_i = X_i + D_i \tag{1.2}$$

Note, however, that the distributions for X_i and D_i are in general correlated,

$$\mathrm{Cov}(X_i, D_i) = \mathrm{Var}(X_i) - \mathrm{Cov}(X_i, Y_i) \tag{1.3}$$

If the underlying distribution was a bivariate Gaussian, we'd expect the data to be spread out roughly in an ellipse in the x-y plane:

*Copyright 2019, John T. Whelan, and all that



Note that this has a lot more structure than is reflected in just $\mu_D = \mu_1 - \mu_2$.

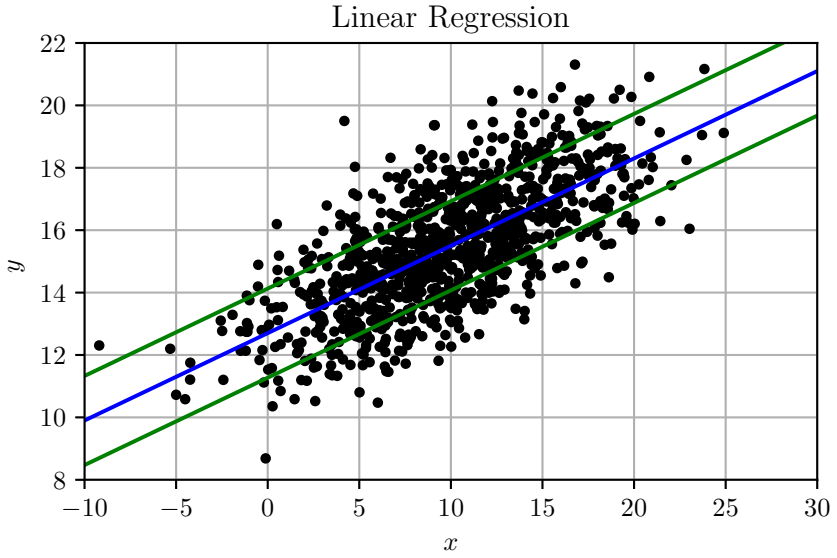
In the case of linear regression, we want to consider data which are once again paired into x_i, y_i, but now the model we assume is that y_i is a linear function of x_i plus some error whose properties don't depend on x_i. In terms of random variables, this means

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1.4}$$

where β_0 and β_1 are constants, and ϵ_i follows a distribution which is independent of X_i; for concreteness we will assume $\epsilon_i \sim N(0, \sigma^2)$ where the error amplitude σ is some non-negative constant. It is also conventional to assume that the values $\{x_i\}$ are known constants, and thus write

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{1.5}$$

although the data analysis is the same in any event. Data associated with a linear regression model are typically spread out about a diagonal line, with the typical offset values being the same all along the line:



1.1 Estimation of Parameters

The first task to be performed in a linear regression problem is to estimate the parameters β_0 and β_1, as well as the error amplitude σ. The usual method for the slope β_1 and intercept β_0 is known as least squares, which finds the values which minimize

$$q(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{1.6}$$

This seems a bit ad hoc at first, and in fact it appears unconnected to our previous method because the techniques of regression and least squares estimation predate the development of classical statistics. But we can actually motivate the process by

a standard technique: maximum likelihood estimation. For a given x_i, $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, which means the pdf is

$$f_i(y_i; \beta_0, \beta_1, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right) \tag{1.7}$$

and the log-likelihood is

$$\begin{aligned} \ln f(y_1, \ldots, y_n; \beta_0, \beta_1, \sigma) &= \ln \prod_{i=1}^{n} f_i(y_i) = \sum_{i=1}^{n} \ln f_i(y_i) \\ &= -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \end{aligned} \tag{1.8}$$

For any value of σ, then, the β_0 and β_1 values which maximize the likelihood will be those which minimize $q(\beta_0, \beta_1)$. The partial derivatives are

$$\frac{\partial q}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \tag{1.9a}$$

$$\frac{\partial q}{\partial \beta_1} = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) \tag{1.9b}$$

and setting both to zero gives the system of equations for the maximum likelihood estimates $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$n\hat{\beta}_0 + \left(\sum_{i=1}^{n} x_i\right)\hat{\beta}_1 = \sum_{i=1}^{n} y_i \tag{1.10a}$$

$$\left(\sum_{i=1}^{n} x_i\right)\hat{\beta}_0 + \left(\sum_{i=1}^{n} x_i^2\right)\hat{\beta}_1 = \sum_{i=1}^{n} x_i y_i \tag{1.10b}$$

which can also be written

$$\hat{\beta}_0 + \overline{x}\hat{\beta}_1 = \overline{y} \tag{1.11a}$$

$$\overline{x}\hat{\beta}_0 + \overline{x^2}\hat{\beta}_1 = \overline{xy} \tag{1.11b}$$

and has solutions

$$\hat{\beta}_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - (\overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{1.12a}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} \tag{1.12b}$$

Note that the second expression for $\hat{\beta}_1$ is less prone to round-off errors, but the first can be easily computed using the first and second moments of the data (n, $\sum_{i=1}^{n} x_i$, $\sum_{i=1}^{n} y_i$, $\sum_{i=1}^{n} x_i^2$, $\sum_{i=1}^{n} y_i^2$, and $\sum_{i=1}^{n} x_i y_i$) as long as extra precision is kept in intermediate steps. However, a more robust (but equivalent) set of summary variables is n, $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$, $S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$, and , $S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2$. In particular, we have

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{1.13a}$$

$$\hat{\beta}_0 = \overline{y} - \frac{S_{xy}}{S_{xx}}\overline{x} \tag{1.13b}$$

which is also fairly easy to remember.

For any x, we can generate a point estimate of the corresponding y, i.e., $\hat{\beta}_0 + \hat{\beta}_1 x$. In particular, for x_i the corresponding "predicted" value is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, and we call $y_i - \hat{y}_i$ the *residual*.

We can return now to the question of estimating σ. One obvious point estimate would be the maximum likelihood value; setting the partial derivative of (1.8)

$$\frac{\partial \ln f}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \tag{1.14}$$

to zero would give a maximum likelihood estimate of σ^2 equal to

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{1.15}$$

however, as with the case for the mle of σ^2 from a normal random sample, this turns out to be a biased estimator because we have to estimate the two parameters β_0 and β_1 from the data. The unbiased estimator is

$$\hat{\sigma}^2 = s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{\text{SSE}}{n-2} \tag{1.16}$$

where we have defined the sum square error (SSE) associated with the estimates $\{\hat{y}_i\}$. A little bit of algebraic manipulation shows that we can also calculate this as either

$$\text{SSE} = \sum_{i=1}^{n} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i \tag{1.17}$$

or

$$\text{SSE} = S_{yy} - \frac{S_{xy}^2}{S_{xx}} \tag{1.18}$$

depending on which set of summary variables we want to work with.

Sometimes it's useful to quantify how much of the variability of the $\{y_i\}$ is explained by the best-fitting linear model. We note that $\text{SSE} = q(\hat{\beta}_0, \hat{\beta}_1)$ is the lowest possible value of $q(\beta_0, \beta_1)$. If we instead restricted ourselves to models with $\beta_1 = 0$, i.e., assumed that $\{y_i\}$ was a sample from a normal distribution $N(\beta_0, \sigma^2)$, we would find that the best fitting value for β_0 was $\overline{y}$ and therefore the lowest value of $q(\beta_0, 0)$ is

$$q(\overline{y}, 0) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \text{SST} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 \tag{1.19}$$

This is called the SST[1] or total sum of squares, and it is a measure of the deviation of the points from the best-fitting horizontal line (whereas the SSE is the deviation from the best-fitting diagonal line).[2] Since $q(\beta_0, \beta_1)$ as defined is always non-negative, and the SSE is the lowest possible value of the function, we have

$$0 \leq \text{SSE} \leq \text{SST} \tag{1.20}$$

The ratio SSE/SST can thus be used to measure how important it was to use a diagonal rather than a horizontal line for the fit. If it's close to 1, then a horizontal line fits almost as well; if it's close to 0, then the diagonal line gives a much better fit. We define the *coefficient of determination*

$$r^2 = 1 - \frac{\text{SSE}}{\text{SST}} \tag{1.21}$$

as a measure ($0 \leq r^2 \leq 1$) of how much better the diagonal line fits than the horizontal line.

Practice Problems

12.9, 12.11, 12.17, 12.19

[1] Note that $\text{SST} = S_{yy}$.

[2] The analogy to the corresponding quantities from ANOVA is that the SSE is the deviation from the best estimate using the model (either that each treatment has its own mean in the case of ANOVA, or that the mean is linearly related to x in the case of regression, whereas the SST is the deviation from the overall average of all of the data, ignoring the model (either treatments or the value of x).

Tuesday 28 March 2019

1.1.1 Estimators in Terms of Summary Data

All of the estimators so far:

$$\hat{\beta}_0 = \overline{y} - \frac{S_{xy}}{S_{xx}}\overline{x} \tag{1.22a}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{1.22b}$$

$$s^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) \tag{1.22c}$$

1.2 Inference for the Slope Parameter

So far, we've considered point estimates for the parameters β_0, β_1 and σ^2, which were

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \tag{1.23}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \tag{1.24}$$

and

$$s^2 = \frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{1.25}$$

respectively. Now we turn to interval estimation and hypothesis testing. We'll focus on the slope parameter β_1, and consider the statistical properties of the estimator

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \tag{1.26}$$

Note that since we're treating the $\{x_i\}$ as given, the only randomness associated with the estimator comes from the $\{Y_i\}$.

Also, note that

$$\sum_{i=1}^n (x_i - \overline{x})\overline{Y} = \overline{Y}\left(\sum_{i=1}^n x_i - n\overline{x}\right) = 0 \tag{1.27}$$

so

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})Y_i}{S_{xx}} \tag{1.28}$$

where $S_{xx} = \sum_{i=1}^n (x_i - \overline{x})^2$. (Note that despite being written with a capital letter, S_{xx} is *not* a statistic or other random variable.) Since $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, $\hat{\beta}_1$, which is a linear combination of the independent normal random variables $\{Y_i\}$, is normal with

$$E(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \overline{x})(\beta_0 + \beta_1 x_i)}{S_{xx}} = \beta_1 \tag{1.29}$$

and

$$V(\hat{\beta}_1) = \frac{\sum_{i=1}^n (x_i - \overline{x})^2 \sigma^2}{(S_{xx})^2} = \frac{\sigma^2}{S_{xx}} = \sigma^2_{\hat{\beta}_1} \tag{1.30}$$

To make a standardized statistic, we have to estimate $\sigma_{\hat{\beta}_1} = \sigma/\sqrt{S_{xx}}$ using the estimate s in place of σ. Specifically,

$$S^2_{\hat{\beta}_1} = \frac{S^2}{S_{xx}} = \frac{1}{(n-2)S_{xx}}\sum_{i=1}^n (Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{1.31}$$

Since the variance estimate required the estimation of two parameters, the standardized statistic

$$T = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{S^2/S_{xx}}} \tag{1.32}$$

obeys a Student t distribution with $n-2$ degrees of freedom.

We can apply this as usual; as a pivot variable, it gives us a confidence interval at CL $(1-\alpha)\times 100\%$ with endpoints

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2} s/\sqrt{S_{xx}} \tag{1.33}$$

Similarly, if we want to test a hypothesis H_0 which says $\beta_1 = \beta_1^{(0)}$ for some proposed value $\beta_1^{(0)}$, the test statistic is

$$t = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{s/\sqrt{S_{xx}}} \tag{1.34}$$

which we compare to $t_{\alpha,n-2}$, $-t_{\alpha,n-2}$ or $\pm t_{\alpha/2,n-2}$, depending on whether it's appropriate to carry out an upper-, lower-, or two-tailed test.

Practice Problems

12.31, 12.79

Tuesday 2 April 2019

Review for Prelim Exam Two (from section 9.1 to 12.1, inclusive). Please bring questions, and ideally ask them by email before class.

Thursday 4 April 2019

Prelim Exam Two (from section 9.1 to 12.1, inclusive). Closed book, closed notes, but you may bring one handwritten 8.5"×11" (front and back) formula sheet, and also use a scientific calculator.

Tuesday 9 April 2019

1.3 Inferences about Predicted Values

1.3.1 Confidence Interval for Average Model Value

Reminder of the linear regression model: Treating the $\{x_i\}$ as given, the observed data are independent rvs, with

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \tag{1.35}$$

where β_0. β_1 and σ are treated as unknown variables. The maximum likelihood/least squares estimators of β_1 and β_0 are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{1}{S_{xx}} \sum_{i=1}^{n} (x_i - \overline{x}) Y_i \tag{1.36}$$

and

$$\hat{\beta}_0 = \overline{Y} - \overline{x}\hat{\beta}_1 \tag{1.37}$$

We've seen (between class and homework) that $\hat{\beta}_0$ and $\hat{\beta}_1$ are each normally distributed unbiased estimators of the corresponding parameter. But rather than just placing confidence intervals on the slope and intercept of the linear function giving the expectation value of the data as a function of x, we're often interested in the uncertainty of where that line is at each point. I.e., we'd like to estimate the quantity $\beta_0 + \beta_1 x$ for a given x. Devore refers to the x value of interest as x^*, and the quantity of interest as

$$\mu_{Y \cdot x^*} = \beta_0 + \beta_1 x^* \tag{1.38}$$

The obvious estimator is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* = \overline{Y} + \hat{\beta}_1 (x^* - \overline{x}) = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x^* - \overline{x})(x_i - \overline{x})}{S_{xx}} \right) Y_i \tag{1.39}$$

which Devore calls $\hat{Y}$. Since it's a linear combination of the $\{Y_i\}$, it's a normally-distributed random variable. As a statistic, it has expectation value

$$E(\hat{Y}) = E(\hat{\beta}_0) + E(\hat{\beta}_1)x^* = \beta_0 + \beta_1 x^* = \mu_{Y \cdot x^*} \tag{1.40}$$

I.e., it's an unbiased estimator. Since the $\{Y_i\}$ are independent random variables, the variance is

$$V(\hat{Y}) = \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(x^* - \overline{x})(x_i - \overline{x})^2}{S_{xx}} \right)^2 V(Y_i) = \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}} \right) \tag{1.41}$$

(We've omitted several steps of non-trivial algebra at the last equals sign!)

As usual, we can't actually calculate $V(\hat{Y})$ from the data; we have to estimate it by replacing σ^2 with

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{1}{n-2} \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right) \tag{1.42}$$

and thus writing

$$s_{\hat{Y}} = s \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}} \tag{1.43}$$

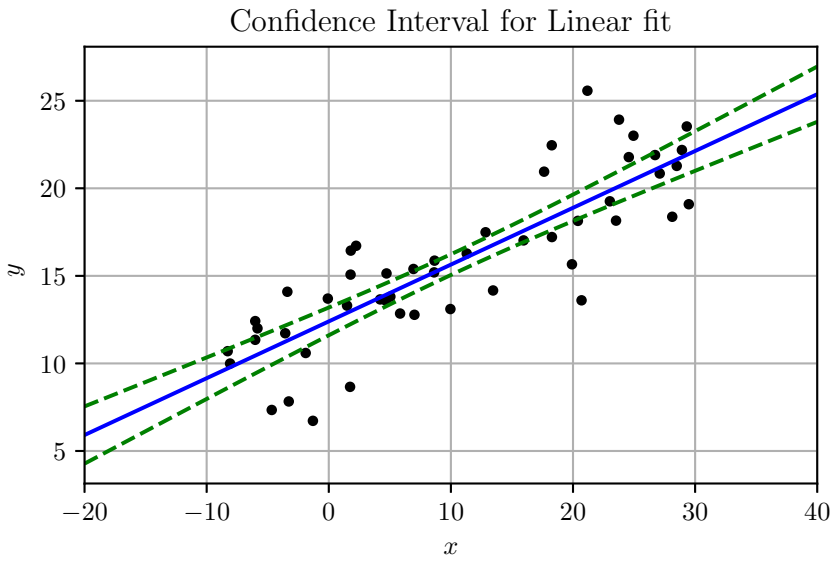
We then know that

$$T = \frac{\hat{Y} - \mu_{Y \cdot x^*}}{S_{\hat{Y}}} \tag{1.44}$$

is a Student-t distributed random variable with $n-2$ degrees of freedom. So for example if we want a confidence interval at confidence level α on $\beta_0 + \beta_1 x^*$ it will be

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2, n-2}\ s \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}} \tag{1.45}$$

Note that the width of this confidence interval depends on x^*. It is a minimum for $x^* = \overline{x}$, and grows as x^* gets farther away from $\overline{x}$. This reflects the fact that, the farther we get from the middle of the data used to estimate the linear relationship between x and y, the less accurately we can estimate the best-fit line.



Note that there are other dangers in trying to extrapolate an inferred linear relationship beyond the range of the observed data. If the linear model we've assumed is not valid over the wider range of x values, we can make a large error which is not reflected in our uncertainty on β_0 and β_1. (Our model doesn't include a β_2, β_3, etc, so we're assuming they're zero, but we don't quantify our uncertainty on that.)

1.3.2 Prediction Intervals for Future Values

Now we wish to consider a slightly different question. Suppose, after making observations of n random variables $\{Y_i\}$ where $Y_i \sim$

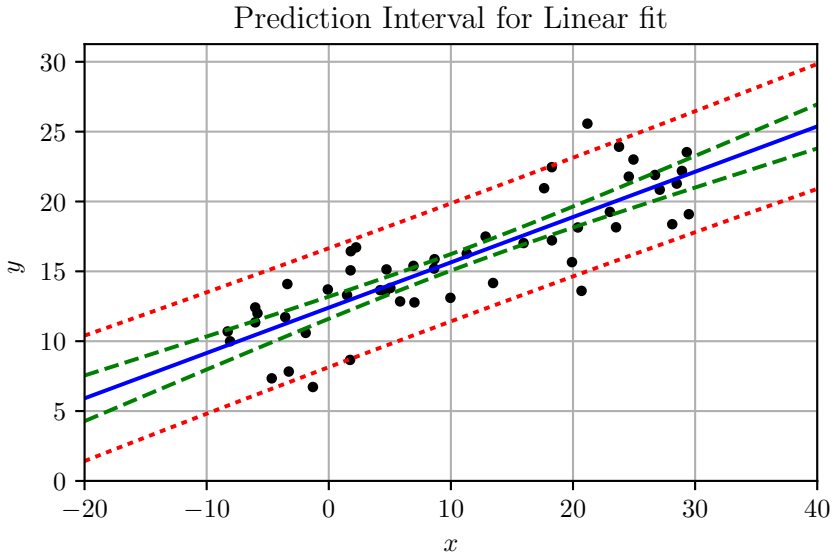
$N(\beta_0+\beta_1 x_i, \sigma^2)$ and doing linear regression estimating using the values $\{y_i\}$, we take a new value Y at x^*. (You could think of this as Y_{n+1} if you like, in which case x^* would be x_{n+1}.) The best estimator we can use for Y, constructed from previous $\{Y_i\}$, is $\hat{Y}$, defined above. Since

$$E(Y) = E(\hat{Y}) = \mu_{Y \cdot x^*} \tag{1.46}$$

the statistic $Y - \hat{Y}$ has zero expectation value. Since $\hat{Y}$ is constructed only from the previous $\{Y_i\}$, Y and $\hat{Y}$ are independent random variables, and thus

$$V(Y-\hat{Y}) = V(Y)+(-1)^2\hat{Y}) = \sigma^2+\sigma^2_{\hat{Y}} = \sigma^2\left(1+\frac{1}{n}+\frac{(x^*-\overline{x})^2}{S_{xx}}\right) \tag{1.47}$$

Thus we see the prediction interval is wider than the confidence interval for the random variable, since it includes the inherent uncertainty associated with each measurement as well as the uncertainty in determining the best-fit line itself:



Practice Problems

12.45, 12.53

Thursday 11 April 2019

1.4 Residuals

(See Devore Section 13.1)

Recall that in our simple linear regression model, we are assuming that Y_1, Y_2, ..., Y_n are independent random variables with

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \tag{1.48}$$

where β_0, β_1 and σ are unknown. The best-fit estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ from the $\{x_i\}$ and the observed $\{y_i\}$. The point estimate of the expected y value at any x is thus $\hat{\beta}_0 + \hat{\beta}_1 x$. In particular, the best-fit model value at x_i is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{1.49}$$

Of course, this is not equal to the actual y_i, so it is useful to plot $e_i = y_i - \hat{y}_i$. According to our statistical model,

$$E(\hat{Y}_i) = E(\hat{\beta}_0) + E(\hat{\beta}_1)x_i = \beta_0 + \beta_1 x_i = E(Y_i) \tag{1.50}$$

so $E(Y_i - \hat{Y}_i) = 0$. We also know $V(Y_i) = \sigma^2$ and, from last time.

$$V(\hat{Y}_i) = \sigma^2\left(\frac{1}{n}+\frac{(x^*-\overline{x})^2}{S_{xx}}\right) \tag{1.51}$$

However, there is one complication in working out $V(E_i) = V(Y_i - \hat{Y}_i)$, which is that, since $\hat{Y}_i$ is constructed using the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, which are in turn constructed from all of the random variables $\{Y_1, Y_2, \ldots, Y_n\}$, the rvs Y_i and $\hat{Y}_i$ are not independent. (Compare this to the situation with prediction intervals, where the value being considered was a new Y (or Y_{n+1},

not included in the original set.) The variance of the residual is thus, when you work out the form of the covariance between

$$\begin{aligned} V(E_i) &= V(Y_i - \hat{Y}_i) = V(Y_i) + V(\hat{Y}_i) - 2\,\mathrm{Cov}(Y_i, \hat{Y}_i) \\ &= \sigma^2 \left(1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{S_{xx}}\right) \end{aligned} \tag{1.52}$$

Note that this is smaller than σ^2, and gets *smaller* when x_i is farther from the mean of the xs. This is because the data themselves are pulling the best fit line towards them, and so even if there are big fluctuations, those fluctuations will influence the best-fit model.

One can plot the *standardized residuals*

$$e_i^* = \frac{y_i - \hat{y}_i}{s\sqrt{1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{S_{xx}}}} \tag{1.53}$$

and see if they are indeed Student-t distributed with $n-2$ degrees of freedom.

2 Correlation

(See Devore Section 12.5)

Finally, we consider a situation where the picture, at least in terms of classical statistics, is different from the regression formalism, but the formulas are related. We've been assuming the $\{x_i\}$ were just given numbers and the $\{Y_i\}$ were random variables with realizations $\{y_i\}$. We could also imagine that the $\{x_i\}$ themselves were the realizations of random variables $\{X_i\}$.[3]

[3]This is largely a "distinction without a difference" because the standard regression treatment works if you consider all of the distributions for Y_i to be conditional distributions for a given realization of $\{x_i\}$, i.e., $f_Y(y_i)$ is actually $f_{Y|X}(y_i|x_i)$.

The rvs are not all independent, but we assume that $\{X_i, Y_i\}$ is a sample drawn from a bivariate distribution, which means that (if the distribution is continuous) the joint pdf is

$$f(x_1, \ldots, x_n, y_1, \ldots, y_n) = f(x_1, y_1)\, f(x_2, y_2) \cdots f(x_n, y_n) \tag{2.1}$$

We will now see that a quantity defined in the context of regression actually has a familiar interpretation when considered in the bivariate context.

Recall the coëfficient of determination r^2, defined as

$$\begin{aligned} r^2 &= 1 - \frac{\mathrm{SSE}}{\mathrm{SST}} = 1 - \frac{1}{S_{yy}}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) = \frac{S_{xx}S_{yy} - S_{xx}S_{yy} + S_{xy}^2}{S_{xx}S_{yy}} \\ &= \frac{S_{xy}^2}{S_{xx}S_{yy}} \end{aligned} \tag{2.2}$$

so that $0 \le r^2 \le 1$. Now, S_{xx} and S_{yy} are both positive, but S_{xy} need not be. So we can define

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \tag{2.3}$$

This is called the *sample correlation coëfficient*, and obviously the square of r is the coëfficient of determination, which we've already called r^2. If we recall the definition of the sample variances

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{S_{xx}}{n-1} \tag{2.4a}$$

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{S_{yy}}{n-1} \tag{2.4b}$$

we see that

$$r s_x s_y = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{S_{xy}}{n-1} \tag{2.5}$$

This bears a striking similarity to the formulas for the variance, covariance and correlation of a pair of random variables:

$$\sigma_x^2 = V(X) = E([X - \mu_X]^2) \tag{2.6a}$$
$$\sigma_y^2 = V(Y) = E([Y - \mu_Y]^2) \tag{2.6b}$$
$$\rho\sigma_x\sigma_y = \mathrm{Cov}(X, Y) = E([X - \mu_X][Y - \mu_Y]) \tag{2.6c}$$

So it should not be too surprising that if we have a sample $\{x_i, y_i\}$ drawn from a bivariate distribution $f(x, y)$ or $p(x, y)$, the sample correlation coëfficient r will be an estimator for the population correlation coëfficient ρ. In terms of a random sample, we can refer to the statistic

$$R = \frac{\sum_{k=1}^{n}(X_k - \overline{X})(Y_k - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{j=1}^{n}(Y_j - \overline{Y})^2}} \tag{2.7}$$

where we've given the sum indices different labels to stress that there are three separate sums in the definition (not counting those used to define $\overline{X}$ and $\overline{Y}$).

The sample correlation coëfficient can be used as a statistic to perform inferences about the correlation present in the population. For simplicity, we'll consider only the case of a hypothesis test where the null hypothesis is $\rho = 0$. In the case where the underlying distributions are both normal, symmetry tells us that $E(R) = 0$. It turns out (you'll sort of show this on the homework) that

$$T = \frac{R\sqrt{n-2}}{\sqrt{1 - R^2}} \tag{2.8}$$

obeys a Student-t distribution with $n - 2$ degrees of freedom.

Practice Problems

12.59, 12.83