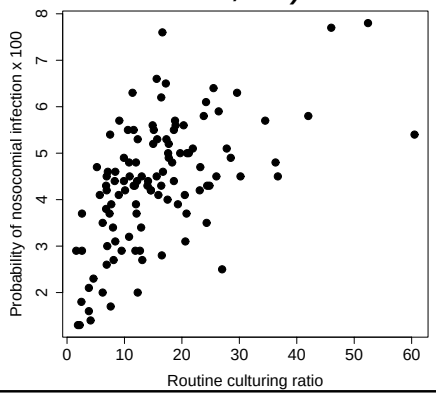
Simple linear regression

Biometry 755

Spring 2008

Overview of regression analysis

- Evaluate relationship between one or more independent variables ($X_1, \ldots, X_k$) and a single *continuous* dependent variable (Y).
- Terminology: "Regress Y on X"
- Example Relationship between risk of nosocomial infection (dependent variable, Y) and routine culturing ratio (independent variable, X)



Goals of a regression analysis

1. Characterize direction and strength of relationship.
2. Prediction
3. Control for effects of other variables
4. Identify group of independent variables that collectively describe the structure (explain the variability) in a random sample of dependent measures.

Goals of a regression analysis (cont.)

5. Describe the best mathematical model for describing relationship between dependent and independent variables.
6. Comparison of associations for two groups
7. Assess interactive effects of two independent variables.
8. Obtain precise estimates of regression coefficients (especially β_1)

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Review of simple linear regression (SLR)

- One dependent variable, Y, and one independent variable, X.
- Observed data consists of n pairs of observations, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- Assume linear function describes the relationship between Xs and Ys.
- Linear function takes the form

$$y_i = \underbrace{\underbrace{\beta_0}_{\text{intercept}} + \underbrace{\beta_1}_{\text{slope}} x_i}_{\text{signal}} + \underbrace{\varepsilon_i}_{\text{noise}}$$

Components of the SLR model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

y_i	Value of dependent variable for ith subject
β_0	True y-intercept.
β_1	True slope.
x_i	Value of independent variable for ith subject.
ε_i	Random error associated with the ith subject. Assumed to be zero, on average. Accounts for 'spread' of the data points around the line.

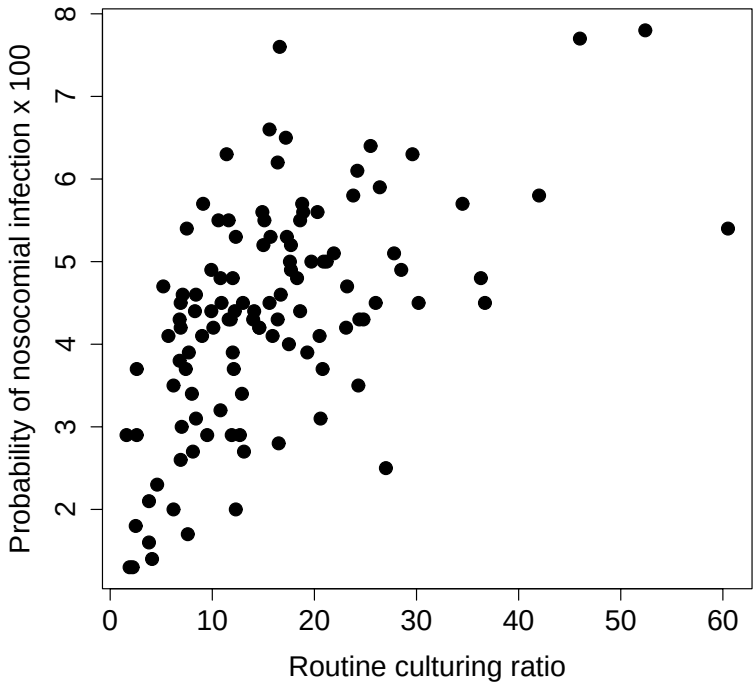
Estimating the components of the SLR model

The true y-intercept and the slope, β_0 and β_1, are unknown.

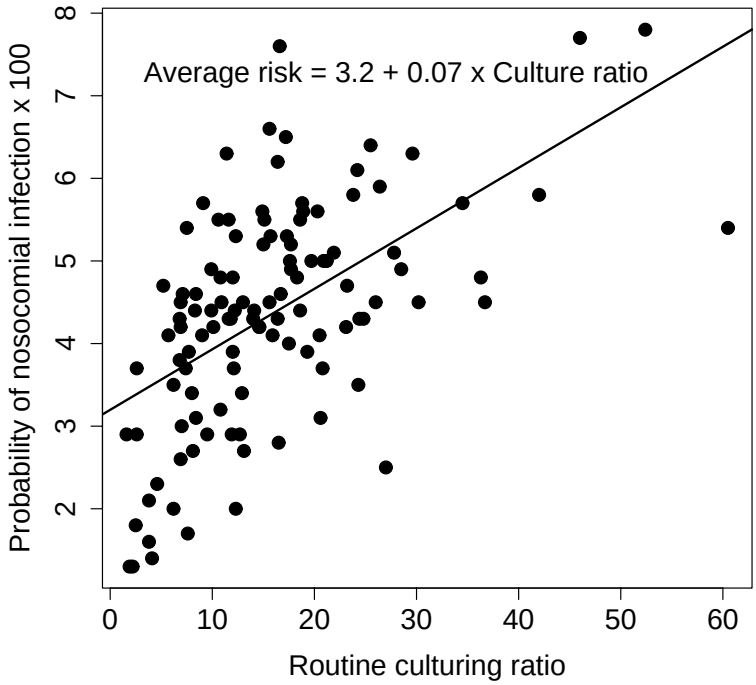
The best we can do is estimate their values, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively, using a method deemed optimal. In SLR, the optimal technique is called the method of *least-squares*. The least-squares estimates of β_0 and β_1 (and hence the least-squares estimate of the line itself) are those that minimize the sum of the squared deviations between the observed data points and the estimated (fitted) line.

For SLR, the least-squares estimates are optimal in the sense that $\hat{\beta}_0$ and $\hat{\beta}_1$ have minimal variance (good precision) and are unbiased (provided the model is correct ... a very BIG assumption!).

Find the 'best-fitting' line



Here is the ‘best-fitting’ line



Least squares estimates of β_0 and β_1

It can be shown that the expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of the squared deviations around the regression line are:

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

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Least squares estimates of β_0 and β_1 (cont.)

Once we have $\hat{\beta}_0$ and $\hat{\beta}_1$, we can estimate the response at x_i based on the fitted regression line. We denote this estimated response as $\hat{y}_i$ and write

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

Note that there is no "$\hat{\varepsilon}_i$" in the expression for the fitted regression line. This is due to the fact that the error around the line is assumed to be zero, on average, and therefore does not contribute to the 'signal' or 'structure' in the data.

SLR in SAS

```
proc reg data = one;
    model infrisk = cult;
run;
```

The regression line is the *average* response

For the SLR model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the average value of y_i given x_i, and its estimate, are represented as ...

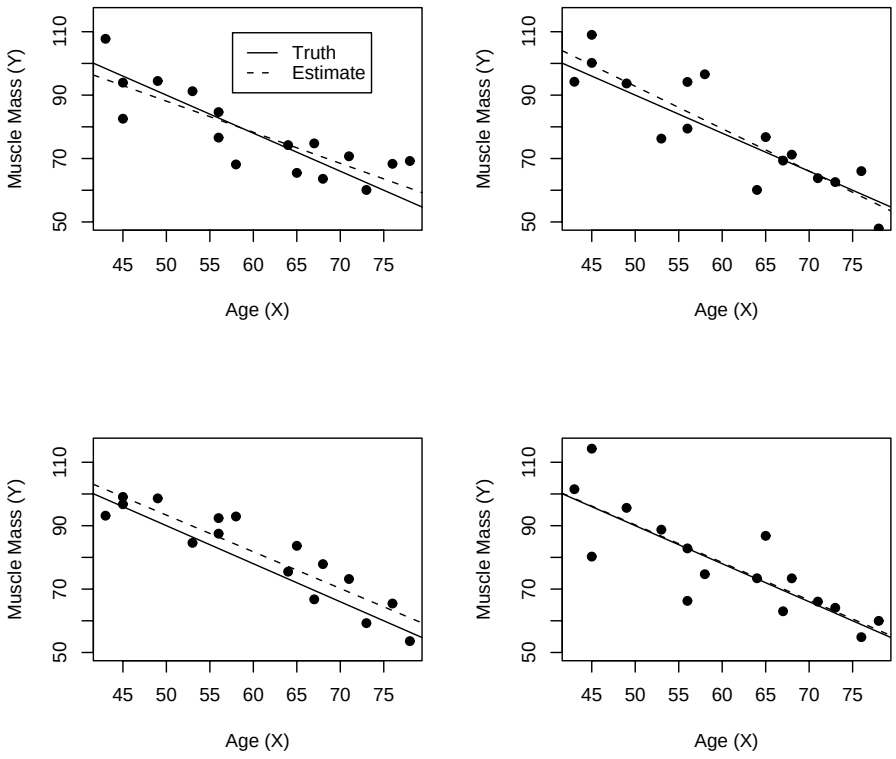
Truth		Estimate
$E(y_i \mid x_i) = \beta_0 + \beta_1 x_i$		$\widehat{E(y_i \mid x_i)} = \hat{\beta}_0 + \hat{\beta}_1 x_i$
	or	
		$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

The estimated average varies

In each of the following panels, a sample of sixteen data points was selected from the same underlying linear model with the same spread about the line, or error variance (σ^2). But each sample of points is different (referred to as *sampling variability*). Therefore the model fit to the sampled points in each example is different, despite the fact that the same true model generated all four different fitted models.

True average given x_i	Solid line	$E(y_i \mid x_i) = \beta_0 + \beta_1 x_i$
Data generated from model	Data points	$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
Estimated average given x_i	Dotted line	$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

The estimated average varies



SLR statistical assumptions

Linearity Given any value of x, y is on average a straight-line function of x. We write

$$E(y|x) = \beta_0 + \beta_1 x$$

where the notation $E(y|x)$ is interpreted in words as "the average value of y given x".

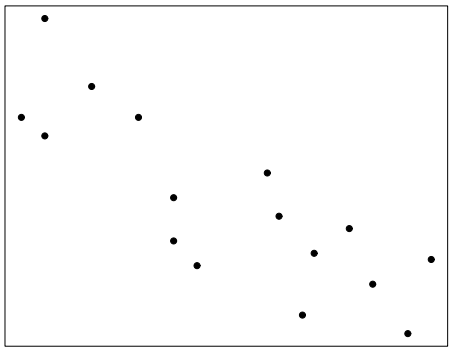
Independence The y_is are statistically independent, i.e. represent a random sample from the population.

Homoscedasticity The variance of y is the same, regardless of the value of x. The variance of y is denoted in the usual manner as σ^2.

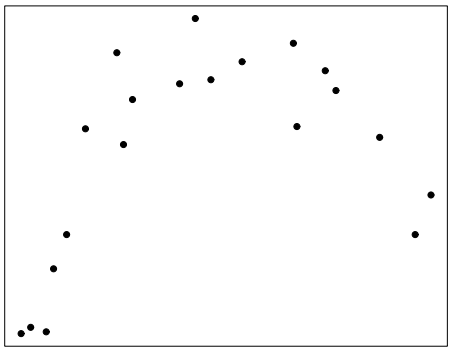
Normality For each value of x, $y \sim \text{Normal}(\beta_0 + \beta_1 x, \sigma^2)$.

Visualizing SLR assumptions

Linearity



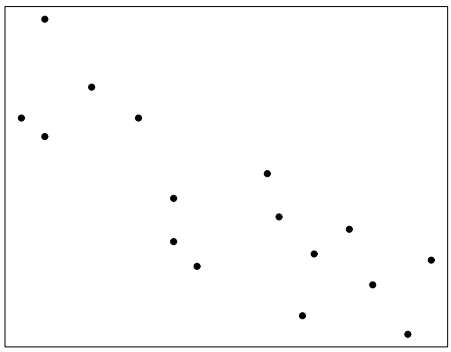
Straight−line model appropriate



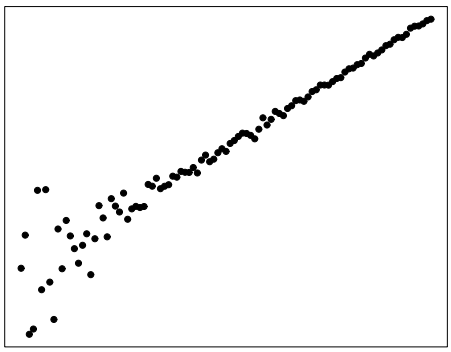
Curvilinear model appropriate

Visualizing SLR assumptions (cont.)

Homoscedasticity



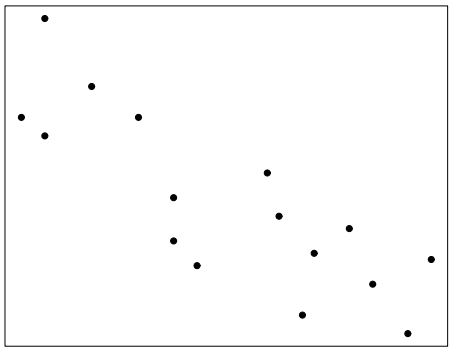
Constant 'spread'



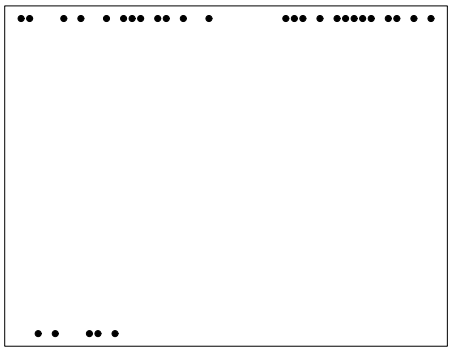
Non−constant 'spread'

Visualizing SLR assumptions (cont.)

Normality



Normal data

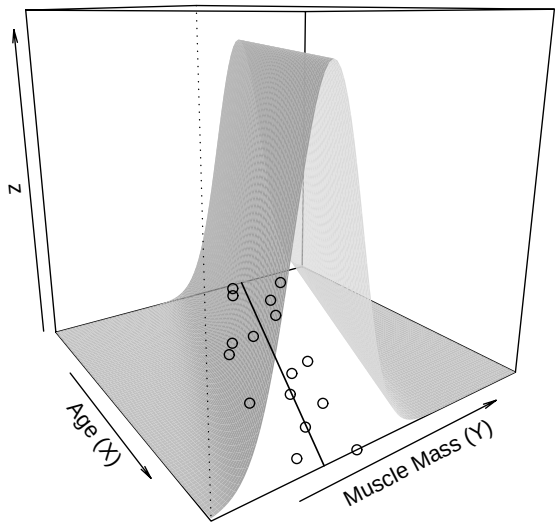


Non−normal data

"That looks normal?!!?"

Visualizing SLR assumptions (cont.)

Normality **(cont.)**



Model errors (ε_is) and residuals (r_is)

Recall the form of the SLR model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

By some simple algebra, we have

$$\begin{aligned} \varepsilon_i &= y_i - (\beta_0 + \beta_1 x_i) \\ &= y_i - E(y_i|x_i). \end{aligned}$$

This gives us a simple way to estimate the errors in the model. Namely,

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. The $\hat{\varepsilon}_i$s are called *residuals*. We write

$$r_i = \hat{\varepsilon}_i.$$

More comments about errors and residuals

- $y_i \overset{\text{ind}}{\sim} \text{Normal}(\beta_0 + \beta_1 x_i, \sigma^2)$ is equivalent to $\varepsilon_i \overset{\text{ind}}{\sim} \text{Normal}(0, \sigma^2)$.
- *Why?* (Intuitive answer) Recall that $E(y_i|x_i)$, the average value of y_i given x_i, is $\beta_0 + \beta_1 x_i$. Since $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, the average value of the errors must be zero. Otherwise y_i would, on average, be too large (average value of ε_i positive) or too small (average value of ε_i negative).
- The residuals sum to zero.

Residuals in SAS

```
proc reg data = one;
    model infrisk = cult/p;
run;
```


Estimating spread around the regression line

Variance is estimated in the usual manner - that is, sum the squared differences between the observed data points and their fitted values (based on the estimated regression equation), and divide by an appropriate normalizing constant.

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

- s^2 is called the *mean squared error* or MSE.
- Since $y_i - \hat{y}_i = r_i$ (the estimated error terms), $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is called the *sum of squared errors* or SSE.
- $s^2 = \text{MSE} = \frac{\text{SSE}}{n-2}$

Some comments about $\hat{\sigma}^2 = \text{MSE}$

Q *Why do we square the estimated errors?*

A See the last comment on slide 22.

Q *Why do we divide by $n-2$? Doesn't the usual formula for estimated variance divide by $n-1$?*

A We lose 2 degrees of freedom in estimating the mean response in a SLR. The usual formula you learn in introductory statistics reflects the loss of a single degree of freedom in estimating the mean.

Look at SAS output for SLR to see estimated SSE, denominator df, and MSE.

Inference about the slope

The framework for the test of significance for the true slope is $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$. The test statistic is

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}\left(\hat{\beta}_1\right)}$$

- $\text{SE}\left(\hat{\beta}_1\right) = \sqrt{\frac{\text{MSE}}{s_x^2(n-1)}}$ is the standard error of $\hat{\beta}_1$
- $s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ is the sample variance of the x_is
- $t \sim t_{n-2}$ under H_0
- A $(1-\alpha) \times 100$ % confidence interval for β_1 is $\hat{\beta}_1 \pm t_{n-2,1-\alpha/2}\ \text{SE}\left(\hat{\beta}_1\right)$

Inference about the slope (cont.)

P-value and decision rule: Reject the null hypothesis if the p-value is less than α.

Conclusion: If the null is not rejected, this means that we have not found a statistically significant linear relationship between X and Y (at level α). Note that there may be a relationship of some other kind (e.g. a non-linear relationship) between X and Y, so failing to reject the null does not imply that there is *no* relationship.

If the null is rejected, this means that there is a significant linear relationship between X and Y. Note, however, that this does not imply that the linear model is the *best* or most correct way to relate Y to X. The true relationship may be something other than linear or it may include other components in addition to a linear component.

Inference about the intercept

Significance tests on the intercept are rarely performed. The reason is that it is often difficult or impossible (or simply not relevant) to collect sample data at or around $x = 0$. If the sample data does not include values near $x = 0$ (as is most often the case) you cannot trust a hypothesis test that focuses on that region. You do not have adequate data to make good inference.

SLR parameter estimates in SAS

```
proc reg data = one;
    model infrisk = cult/clb;
run;
```


Inference about the regression line

In Slide 15, we saw that different random samples of data points from the same underlying model resulted in different estimated regression lines. This implies that there is some inherent sampling variability associated with the regression line itself. Just as a confidence interval provides information about the precision of point estimate, we would like a two-dimensional equivalent to a CI that provides information about the precision of the fitted regression line. We call such an interval a *confidence band* for the regression line, and it provides an estimate of the uncertainty associated with a SLR.

Inference about the regression line (cont.)

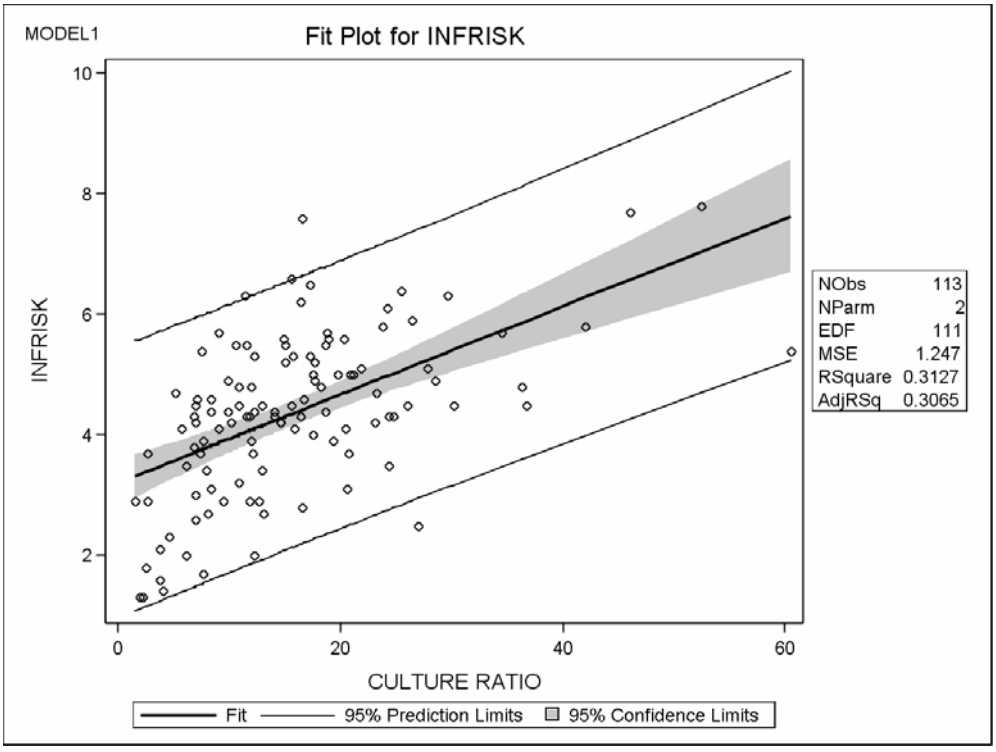
At *each* value of $x = x_0$ (where x_0 is the x-value associated with an observed data point), we compute the $(1-\alpha) \times 100$ % confidence interval

$$\hat{y}|x_0 \pm t_{n-2,1-\alpha/2}\ \mathrm{SE}(\hat{y}|x_0)$$

- $\hat{y}|x_0$ is the fitted value at x_0
- $\mathrm{SE}(\hat{y}|x_0) = s\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{(n-1)s_x^2}}$
- $s = \hat{\sigma} = \sqrt{\mathrm{MSE}}$
- $s_x^2 = \frac{\sum_{i=1}^n (x_i-\bar{x})^2}{n-1}$ is the sample variance of the x_is

At $\alpha = 0.05$, we say that we are 95% confident that the true regression line lies within the confidence band.

Graphing the confidence band



Graphing the confidence band - SAS code

```
ods html;
ods graphics on;
ods select Fit;

proc reg data = one;
    model infrisk = cult/clm;
run;

ods graphics off;
ods html close;
```


Prediction of a new value of y at $x = x_0$

Often the goal of fitting a simple linear regression is to obtain a model that facilitates prediction. In SENIC example, we might be interested in predicting the probability of nosocomial infection at a local hospital where the routine culturing ratio is 30. The obvious estimate is

$$\widehat{\text{Infrisk}} = 3.2 + 0.07 \times 30 = 5.3.$$

Prediction of a new value of y at $x = x_0$ (cont.)

As always, we need a sense of the precision associated with this predicted value. There are two sources of variability associated with predicting the response:

1. The variability associated with the fitted regression line, illustrated in Slide 32. We estimate the square root of this variability with SE$(\hat{y}|x_0)$, defined on Slide 31.
2. The variability associated with data points around the true line, illustrated in Slide 20. We estimate this variability with $\hat{\sigma}^2 = \text{MSE} = s^2$, defined on Slide 24.

Prediction of a new value of y at x_0 (cont.)

The variability associated with predicting a response, y, at $x = x_0$, is the sum of the variability due to fitting the regression line and the variability of the y values at $x = x_0$ around the true regression line. We estimate this total variability as follows:

$$\underbrace{s^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}\right)}_{\text{Var}_{(\hat{y}|x_0)}} + \underbrace{s^2}_{\hat{\sigma}^2} = s^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}\right).$$

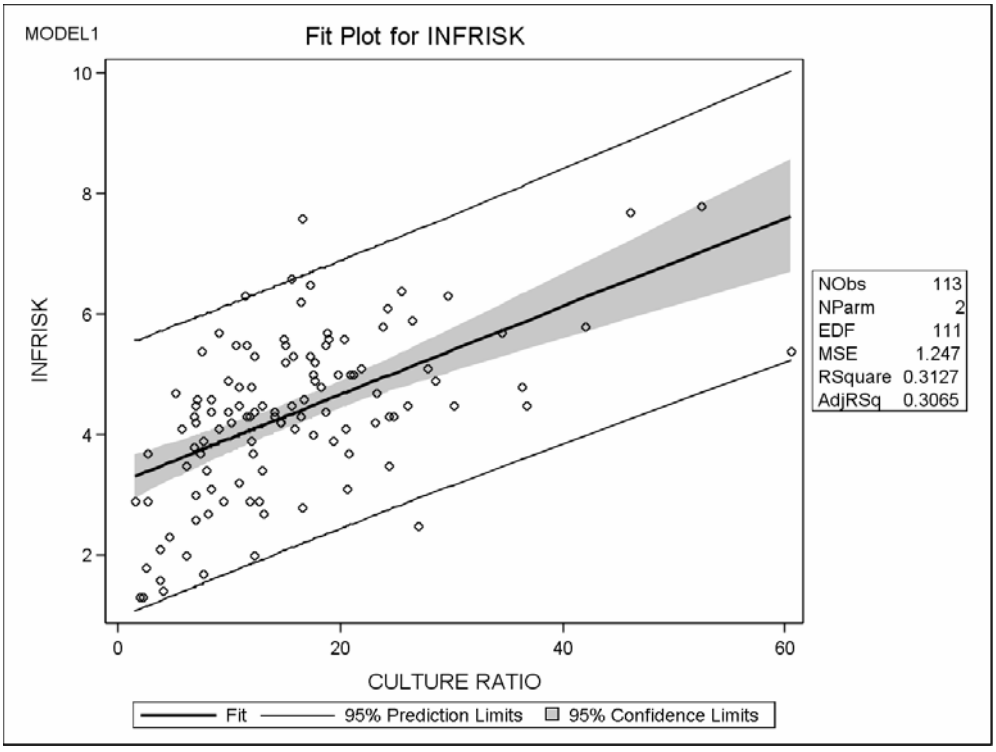
Prediction of a new value of y at x_0 (cont.)

Then a $(1-\alpha)\times 100\%$ prediction interval is constructed using the formula

$$\hat{y}|x_0 \pm t_{n-2,1-\alpha/2}\; s\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{(n-1)s_x^2}}.$$

We say that we are 95% confident that new responses will fall between the upper and lower limits of the interval.

Graphing the prediction interval



Graphing the confidence band - SAS code

```
ods html;
ods graphics on;
ods select Fit;

proc reg data = one;
    model infrisk = cult/cli;
run;

ods graphics off;
ods html close;
```


Interpretation

Confidence interval for mean at $x = x_0$

Prediction interval at $x = x_0$