MA4605 Lecture 4A : Confidence Intervals and Prediction Intervals for Fitted Values

Previously we have seen the confidence intervals for regression coefficients in linear models. These confidence intervals are computed using the ***standard error*** values , which are available on the output of the `summary()` command, when using `R`.

Recall that confidence intervals are generally constructed using point estimates, quantiles and standard errors.

In this class we will look at two more type of intervals.

- Confidence Intervals for Fitted Values
- Prediction Intervals for Fitted Values

Recall: a fitted value ***Y**** is a estimate for the response variable, as determined by a linear model. The difference between the observed value and the corresponding fitted value is known as the residual.

The ***residual standard error*** is the conditional standard deviation of the dependent variable Y given a value of the independent variable X. The calculation of this standard error follows from the definition of the residuals.

$$s_{Y.X} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - 2}} = \sqrt{\frac{\Sigma e^2}{n - 2}}$$

The residual standard error is often called the root mean square error (RMSE), and is a measure of the differences between values predicted by a model or an estimator and the values actually observed from the thing being modelled or estimated.

Since the residual standard error is a good measure of accuracy, it is ideal if it is small.

```
>summary(FitA)
…
…
…
…
Residual standard error: 13.82 on 28 degrees of freedom
Multiple R-squared: 0.302,      Adjusted R-squared: 0.2771
F-statistic: 12.11 on 1 and 28 DF,  p-value: 0.001658
```

(N.B. 30 pairs of covariates were used in ***FitA***, hence 30-2 = 28 degrees of freedom)

This value is used to compute the standard errors for the regression coefficients. For example, standard error for slope in the case of a simple linear regression model is

$$s_{b_1} = \frac{s_{Y.X}}{\sqrt{\Sigma X^2 - n\bar{X}^2}}$$

Confidence Intervals for Fitted Values

Formally this type of confidence interval is formally known as the confidence interval for the conditional mean i.e. E[Y|X]. (For a given value of X, the value that we expect Y to be) .

The point estimate for the conditional mean of the dependent variable Y , given a specific value of X, is the fitted value **Y*.**

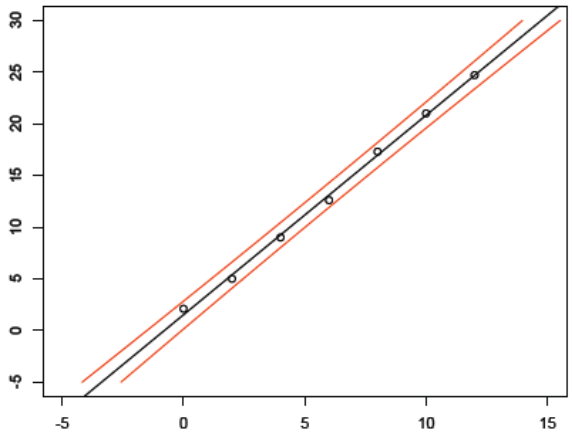
Standard Error of the Conditional Mean (i.e. fitted values) is as follows:

$$s_{\bar{Y}.X} = s_{Y.X}\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma X^2 - [(\Sigma X)^2/n]}}$$

Notice that it uses the definition for ***residual standard error***.

An interesting property of this formula is that the standard errors get narrower, as the value of X gets closer to the mean of X.

Recall the ***Fluoresence v Concentration*** example used in previous classes. The confidence intervals for fitted values are as follows:




```
>Fluo=c(2.1,5.0,9.0,12.6,17.3,21.0,24.7)
>Conc=c(0,2,4,6,8,10,12)

>FitFC=lm(Fluo~Conc)
>FitFC

Call:
lm(formula = Fluo ~ Conc)

Coefficients:
(Intercept)         Conc
      1.518        1.930
```

Recall : the regression equation is

Fluo* = 1.518 + 1.930 Conc

(The asterisk denotes a fitted value, as opposed to an observed value. We use this due to typographical limitations in producing PDFs)

The *R* command we will use is `predict`(). We are going to predict the fluorescence values for new values of concentration (written in the code as ***Conc+1,*** essentially adding 1 to each value of the independent variable). We specify the following:

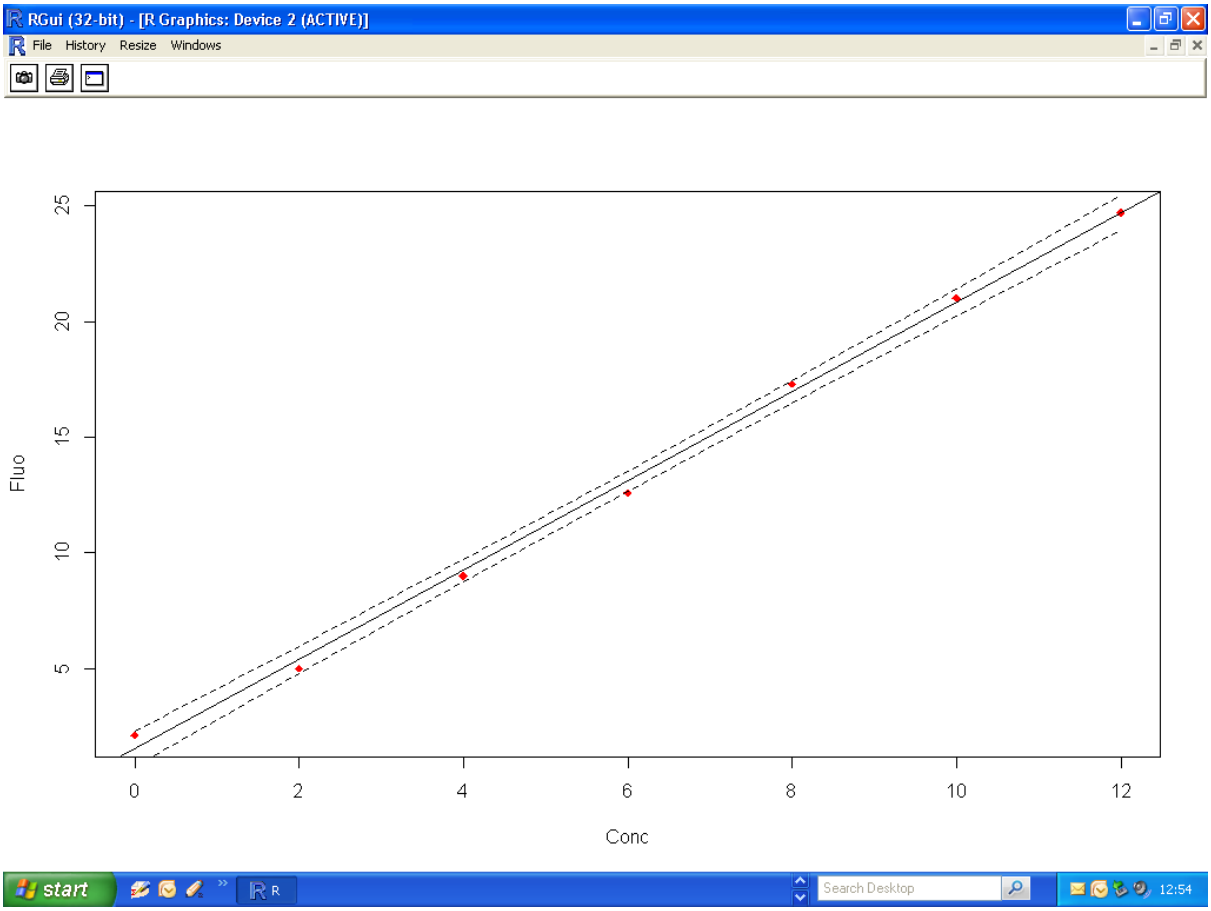
- The name of the fitted model (essentially stating what regression equation to use)
- The new data set (which must be constructed as an object known as "*data.frame*" – we will cover data frames in class soon).
    - If we are using the original data set, we can leave the argument blank.
    - If we are using new data, we must specify the name of the new data as that of the original independent variable.
- We must specify what type of interval we require (i.e. confidence interval or prediction interval)

Let's try this out with the original data first.

```
> predict(FitFC, interval="confidence")
        fit       lwr       upr
1  1.517857  0.759700  2.276014
2  5.378571  4.783824  5.973319
3  9.239286  8.769097  9.709475
4 13.100000 12.679450 13.520550
5 16.960714 16.490525 17.430903
6 20.821429 20.226681 21.416176
7 24.682143 23.923986 25.440300
```

Now let's try this with new data.

```
> Conc
[1]  0  2  4  6  8 10 12
> Conc+1  #The New Data
[1]  1  3  5  7  9 11 13
>
> #Save it as a dataframe
> #Use the same column names as the original Ind. Variables.
> new <- data.frame(Conc=c(Conc+1))
>
>
> predict(lm(Fluo ~ Conc), new, interval="confidence")
        fit       lwr       upr
1  3.448214  2.775006  4.121423
2  7.308929  6.783241  7.834616
3 11.169643 10.736150 11.603136
4 15.030357 14.596864 15.463850
5 18.891071 18.365384 19.416759
6 22.751786 22.078577 23.424994
7 26.612500 25.764855 27.460145
```



This plot was constructed using the following code:

```
> plot(Conc, Fluo, pch =18, col="red")
> abline(coef(lm(Fluo ~ Conc)))
> pred = predict(lm(Fluo ~ Conc),  interval="confidence")
> lines(Conc, pred[,2], lty=2)
> lines(Conc, pred[,3], lty=2)
```

Prediction Intervals

In contrast to a confidence interval, which is concerned with estimating a population parameter, a ***prediction interval*** is concerned with estimating an individual value and is therefore a type of ***probability interval***. (Another type of interval we will see later in the course is a ***Tolerance Interval***)

The complete standard error for a prediction interval is called the ***standard error of forecast***, and it includes the uncertainty associated with the vertical “scatter” about the regression line plus the uncertainty associated with the position of the regression line value itself.

The basic formula for the standard error of forecast is as follows:

$$s_{Y(\text{next})} = \sqrt{s_{Y.X}^2 + s_{\bar{Y}.X}^2}$$

i.e. it is based on the standard error for conditional means.

$$s_{Y(\text{next})} = s_{Y.X}\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\Sigma X^2 - [(\Sigma X)^2/n]}}$$

Finally, the prediction interval for an individual value of the dependent, variable, using ***n - 2*** degrees of freedom, is

$$\hat{Y} \pm ts_{Y(\text{next})}$$

As with the confidence Intervals for fitted values, the *R* command we will use is `predict`().We specify the following:

- The name of the fitted model (essentially stating what regression equation to use)
- The new data set (which must be constructed as an object known as *"data.frame"*)
    - If we are using the original data set, we can leave the argument blank.
    - If we are using new data, we must specify the name of the new data as that of the original independent variable.
- We must specify what type of interval we require (i.e. confidence interval or prediction interval). So here we specify "prediction" rather than "confidence".

```
> predict(lm(Fluo~Conc), new, interval="prediction")
        fit       lwr       upr
1  1.517857  0.759700  2.276014
2  5.378571  4.783824  5.973319
3  9.239286  8.769097  9.709475
4 13.100000 12.679450 13.520550
5 16.960714 16.490525 17.430903
6 20.821429 20.226681 21.416176
7 24.682143 23.923986 25.440300
```

This plot was constructed using the following code, in addition to previous code. A warning error comes when you use the original data.

```
> predfv = predict(lm(Fluo ~ Conc),new, interval="prediction")
> lines(c(Conc+1), predfv[,2], lty=2, col="red")
> lines(c(Conc+1), predfv[,3], lty=2, col="red")
>
>
> predfv = predict(lm(Fluo ~ Conc), interval="prediction")
> lines(Conc, predfv[,2], lty=2, col="red")
> lines(Conc, predfv[,3], lty=2, col="red")
```

Remarks:

A full discussion of the appropriate use of Prediction Intervals, Confidence Intervals and Tolerance Intervals will take place later in the course. Confidence Intervals are well-known, but are often over used or misused, when in fact alternative types of intervals are appropriate.

We will meet Tolerance Intervals in the latter parts of the module, i.e. in the statistical process control section.