PREDICTION AND PREDICTION VARIABILITY

We have seen in lectures that when considering the variability of predictions for new x values, we may wish to account for future residual errors in the calculation. Specifically, our model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

is regarded as relating to **all** outcome data we will observe, whether they be part of the original sample based on predictor values $x_{i1}, i = 1, \ldots, n$ or part of the 'new' sample with predictor values $x_{i1}^{\text{new}}, i = 1, \ldots, m$. We noted the relationship between a predicted value at x_{i1}^{new} ***without*** residual error, $\widehat{Y}_i^{\text{new}}$, and the prediction ***with*** residual error, $\widehat{Y}_{Oi}^{\text{new}}$, as

$$\widehat{Y}_i^{\text{new}} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1}^{\text{new}}$$

with estimators $(\widehat{\beta}_0, \widehat{\beta}_1)$, whereas

$$\begin{aligned} \widehat{Y}_{Oi}^{\text{new}} &= \widehat{Y}_i^{\text{new}} + \epsilon_i^{\text{new}} \\ &= \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1}^{\text{new}} + \epsilon_i^{\text{new}}. \end{aligned}$$

In both cases, the point prediction is simply $\widehat{y}_i^{\text{new}} = \mathbf{x}_i^{\text{new}} \widehat{\beta}$, but the associated uncertainty intervals are different: for a $(1-\alpha)100\%$ interval, we have

$$\begin{aligned} \text{Confidence interval} &: \widehat{y}_i^{\text{new}} \pm t_{\alpha/2, n-2} \sqrt{\left(\widehat{\sigma}^2 \mathbf{H}^{\text{new}}\right)_{ii}} \\ \text{Prediction interval} &: \widehat{y}_i^{\text{new}} \pm t_{\alpha/2, n-2} \sqrt{\widehat{\sigma}^2 \left(\mathbf{I}_m + \mathbf{H}^{\text{new}}\right)_{ii}} \end{aligned}$$

where

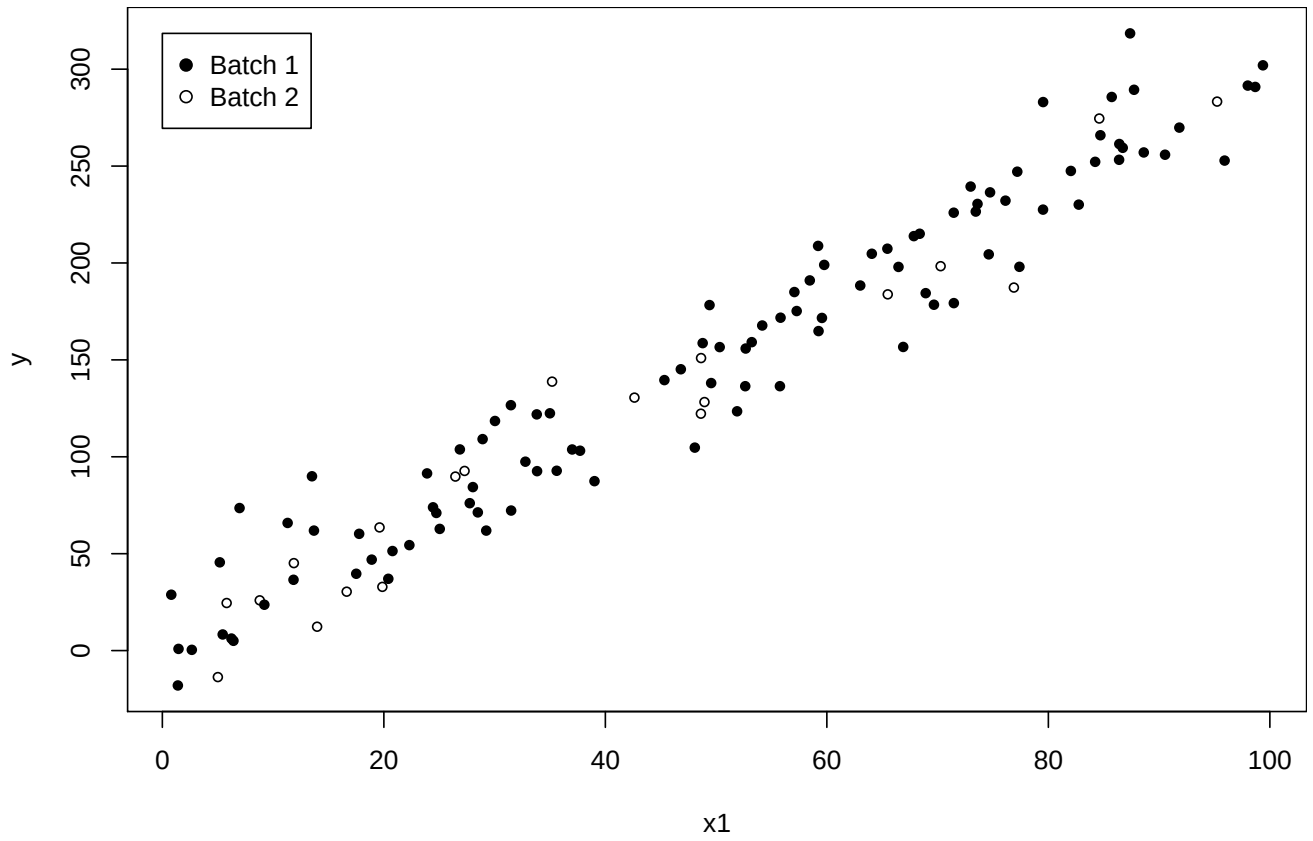
$$\mathbf{H}^{\text{new}} = \mathbf{X}^{\text{new}} (\mathbf{X}^\top \mathbf{X})^{-1} \left\{\mathbf{X}^{\text{new}}\right\}^\top.$$

To illustrate the difference between a confidence interval and a prediction interval, consider the following experiment. We observe $n = 100$ data points in an initial data batch, and then $m = 20$ data points subsequently, with all observations independent. Thus we have a total of 120 observations. The data are simulated using the model

$$Y_i = 2 + 3x_{i1} + \epsilon_i$$

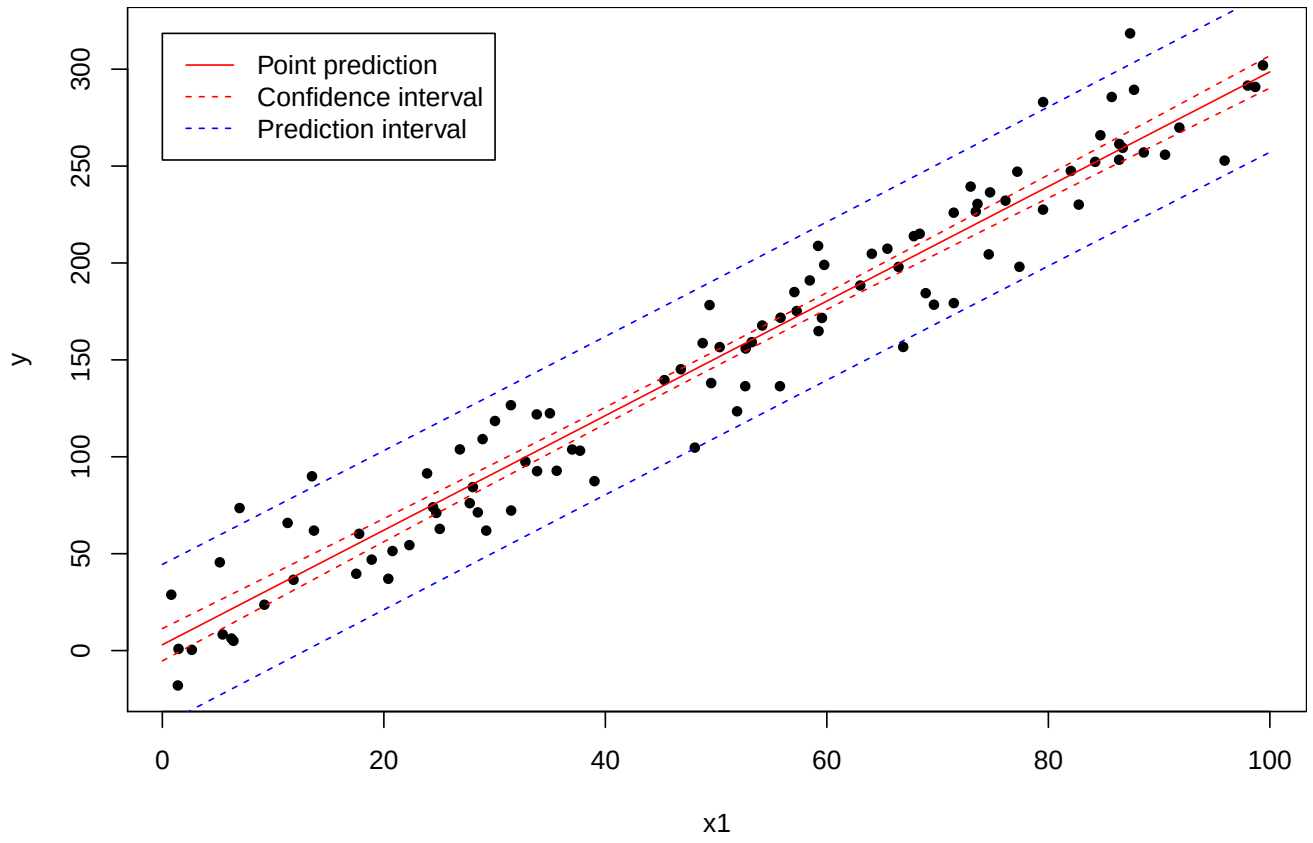
where $\epsilon_i \sim \mathcal{N}(0, 20^2)$.

```
n<-100
m<-20
set.seed(237)
x1<-runif(n,0,100)
y<-2.0+3.0*x1+rnorm(n)*20
x1new<-runif(m,0,100)
ynew<-2.0+3.0*x1new+rnorm(m)*20
par(mar=c(4,4,0,0))
plot(x1,y,pch=19,cex=0.75,xlim=range(c(x1,x1new)),ylim=range(c(y,ynew)))
points(x1new,ynew,pch=1,cex=0.75)
legend(0,max(c(y,ynew)),c('Batch 1','Batch 2'),pch=c(19,1))
```



However, suppose that we fit the model only on the first batch: we may compute the line of best fit, and the predict using a confidence interval and a prediction interval calculation based on the first 100 data points. Here we use $\alpha = 0.05$, and construct 95 % intervals:

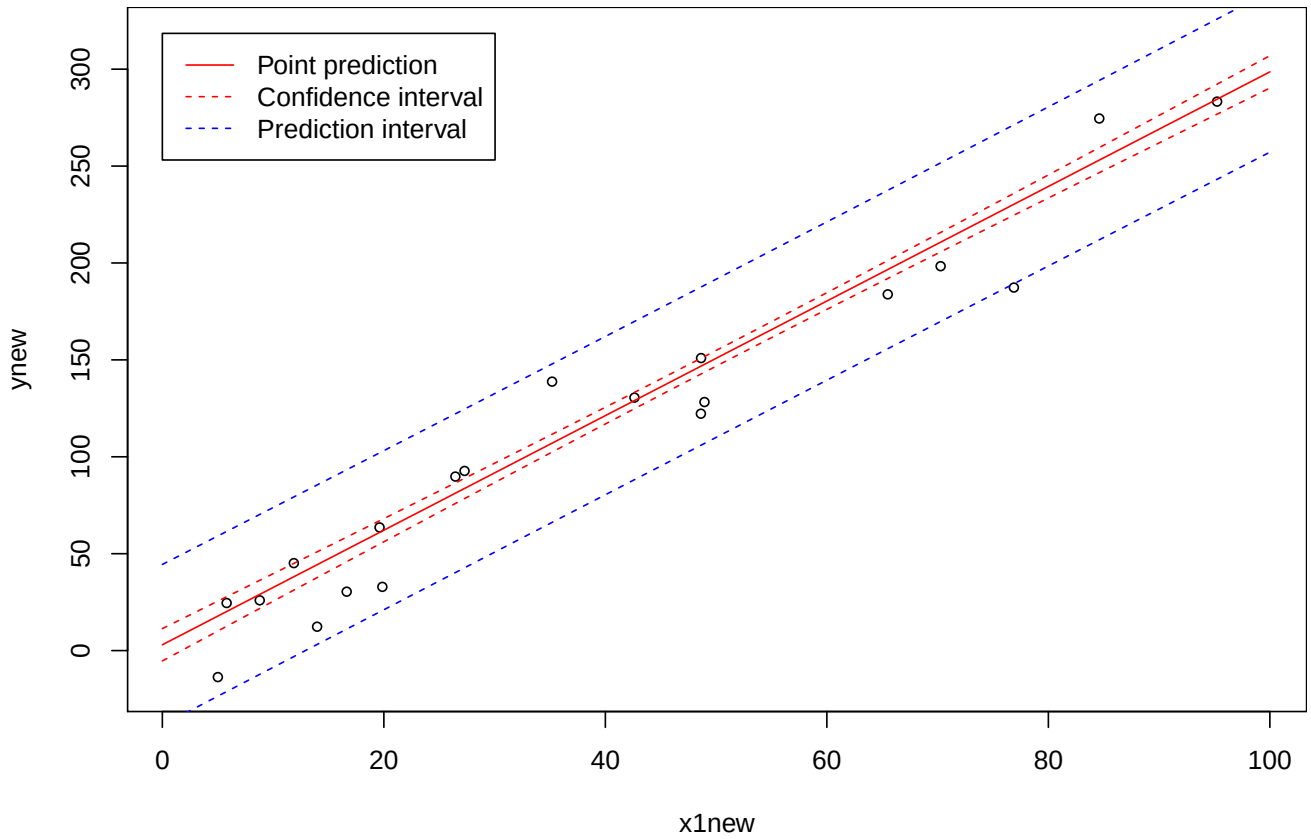
```
fit1<-lm(y~x1)
x1new.vec<-seq(0,100,by=0.1)
conf.interval<-predict(fit1,newdata=data.frame(x1=x1new.vec),interval='confidence')
pred.interval<-predict(fit1,newdata=data.frame(x1=x1new.vec),interval='prediction')
par(mar=c(4,4,0,0))
plot(x1,y,pch=19,cex=0.75,xlim=range(c(x1,x1new)),ylim=range(c(y,ynew)))
lines(x1new.vec,conf.interval[,1],col='red')
lines(x1new.vec,conf.interval[,2],col='red',lty=2)
lines(x1new.vec,conf.interval[,3],col='red',lty=2)
lines(x1new.vec,pred.interval[,2],col='blue',lty=2)
lines(x1new.vec,pred.interval[,3],col='blue',lty=2)
legend(0,max(c(y,ynew)),
       c('Point prediction','Confidence interval','Prediction interval'),
       lty=c(1,2,2),col=c('red','red','blue'))
```




```
conf.interval[1:5,]  #Confidence interval

+        fit       lwr      upr
+ 1 3.002427 -5.341780 11.34663
+ 2 3.298023 -5.033460 11.62951
+ 3 3.593619 -4.725147 11.91238
+ 4 3.889215 -4.416839 12.19527
+ 5 4.184810 -4.108538 12.47816

pred.interval[1:5,]  #Prediction interval

+        fit       lwr      upr
+ 1 3.002427 -38.45273 44.45758
+ 2 3.298023 -38.15457 44.75062
+ 3 3.593619 -37.85642 45.04366
+ 4 3.889215 -37.55828 45.33671
+ 5 4.184810 -37.26014 45.62976
```

The red dashed lines reflect the uncertainty in where the 'true' straight line lies, whereas the blue dashed lines indicate the uncertainty in where future observed responses would lie if a collection of new observations were made. However here, we can compare the intervals with the second batch of observed, but not used, data.

```
par(mar=c(4,4,0,0))
plot(x1new,ynew,pch=1,cex=0.75,xlim=range(c(x1,x1new)),ylim=range(c(y,ynew)))
lines(x1new.vec,conf.interval[,1],col='red')
lines(x1new.vec,conf.interval[,2],col='red',lty=2)
lines(x1new.vec,conf.interval[,3],col='red',lty=2)
lines(x1new.vec,pred.interval[,2],col='blue',lty=2)
lines(x1new.vec,pred.interval[,3],col='blue',lty=2)
legend(0,max(c(y,ynew)),
       c('Point prediction','Confidence interval','Prediction interval'),
       lty=c(1,2,2),col=c('red','red','blue'))
```



Our prediction interval is constructed such that, if the model is correct, 95% of all 'new' observations will lie within the reported interval. In this simulation, with random number generator seed set using the command `set.seed(237)`, 19 out of the 20 new points lie within the interval.