MATH60604A
Statistical modelling
§2f - Predictions

HEC Montréal
Department of Decision Sciences

Predicted and fitted values

- In many applications, particularly in database marketing, the primary goal is to develop a model to get predictions of the dependent variable and use them to make business decisions.
- For example, we might want to predict the amount of money spent if we were to send an offer to a client.
- The usual way of proceeding would be to send offers to a sample of clients, build a model with the data, then apply the model (i.e., obtain predictions) for the other clients in the dataset.

Prediction

- We may want to estimate the mean value of Y when $\mathbf{X} = x$,

$$\mathrm{E}\left(Y \mid \mathbf{X} = x\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

- We also might want to predict the value of a new random variable Y_i when $\mathbf{X}_i = x$; recall that

$$Y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon_i.$$

- Whether we want to predict the mean or the value of Y when $\mathbf{X} = x$, the predicted values will be the point on the "line" corresponding to $\mathbf{X} = x$,

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_p x_p.$$

Individual predictions are more uncertain

- The mean prediction and the prediction for an individual value are the **same** for the linear regression.
  - we will see later that this is not the case for mixed models, in which there are random effects for individuals or groups
- While the estimator of the mean of Y, $\mathsf{E}\left(Y \mid \mathbf{X}\right)$ will be the same, it will be **more precise** than the prediction of an individual value Y_i.
- The **confidence interval** (for the mean) would be smaller than the **prediction interval**.
  - **rationale**: additional uncertainty due to ε for the new observation.
- Once the parameters have been estimated, we can get predicted/fitted values for Y for a given set of predictors $\mathrm{X}_1 = x_1, \ldots, \mathrm{X}_p = x_p$

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_p x_p.$$

Predicted values and prediction intervals in SAS

- We will start by creating a new dataset `newdata` containing the values of the explanatory variables we want to make predictions for.
- We use the first observation and copy all of the explanatories from that individual, but change fixation so that it varies from 0 to 6 seconds.

SAS code to create a new dataset `newdata`

```
data newdata;
   set statmod.intention(obs=1);
   do fixation=0 to 6;
      output;
   end;
run;
```

Saving the output of the linear model for prediction

- Next, we fit the linear model and save the information needed in order to make predictions in an object, say `modelinfo`.
- Next, we use the procedure `plm` to get predictions from the object `modelinfo`. These predictions will be saved in the temporary file `prediction`.
- We also ask for confidence intervals for both the mean (`lclm` and `uclm`) and individual predictions (`lcl` and `ucl`).

SAS code to save fit of linear models

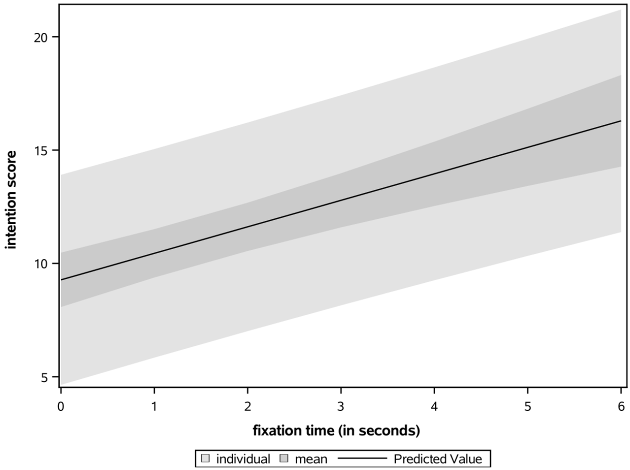
```
proc glm data=statmod.intention noprint;
class sex educ revenue;
model intention= fixation emotion
    sex age revenue educ / ss3 solution;
store modelinfo;
run;
```

SAS code to get predictions using `plm`

```
proc plm restore=modelinfo;
score data=newdata out=prediction predicted
    lclm uclm lcl ucl;
run;

proc sgplot data=prediction;
band x=fixation upper=ucl lower=lcl /
        fill transparency=.5
        legendlabel="individual";
band x=fixation upper=uclm lower=lclm /
        fill transparency=.1 legendlabel="mean";
series x=fixation y=predicted;
yaxis label="intention score";
xaxis label="fixation time (in seconds)";
run;
```

Confidence bands for the mean and prediction bands



The prediction bands (light gray) are wider than the confidence bands for the mean (dark gray). The prediction and confidence intervals get wider when we move away from the mean value of `fixation`