Overfitting and cross validation

Peter Hoff

STAT 423

Applied Regression and Analysis of Variance

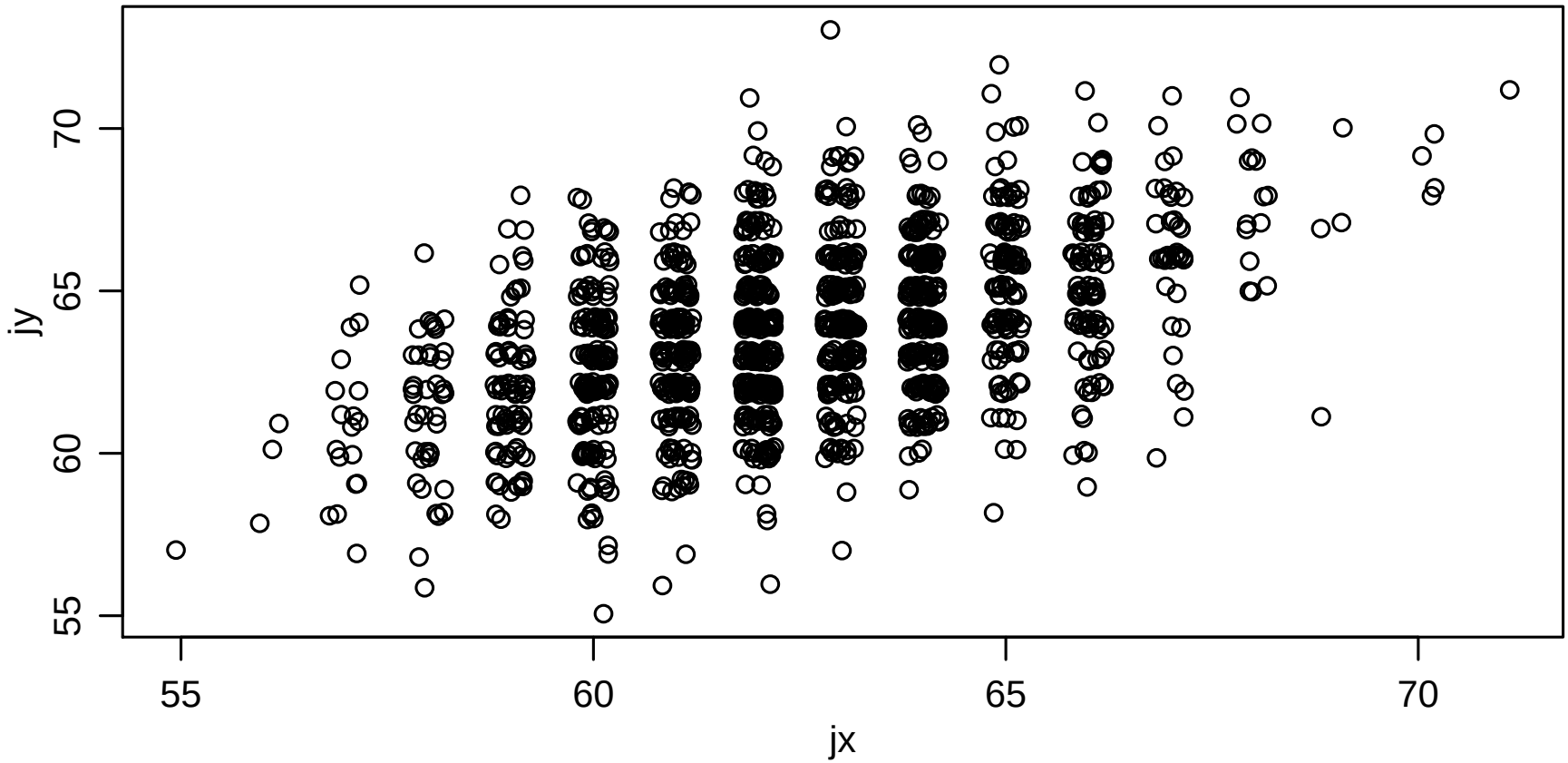
University of Washington

Things to learn

- unrestricted mean model
- linear mean model
- cross validation
- overfitting

Height data

Data on heights of 1375 mother-daughter pairs.



x_i = height of mother i, y_i = height of daughter i

Mean models

Unrestricted model

Mean model: $\mathsf{E}[Y|X = x] = \mu_x,\ x \in \{55, 56, \ldots, 71\}$

Unknown parameters: $\mu_{55}, \mu_{56}, \ldots, \mu_{71}$ (seventeen parameters)

Parameter estimates: $\hat{\mu}_{x'} = \text{mean(y[x==x'])}$

Mean models

Unrestricted model

Mean model: $\mathrm{E}[Y|X = x] = \mu_x,\ x \in \{55, 56, \ldots, 71\}$

Unknown parameters: $\mu_{55}, \mu_{56}, \ldots, \mu_{71}$ (seventeen parameters)

Parameter estimates: $\hat{\mu}_{x'}$ = mean(y[x==x'])

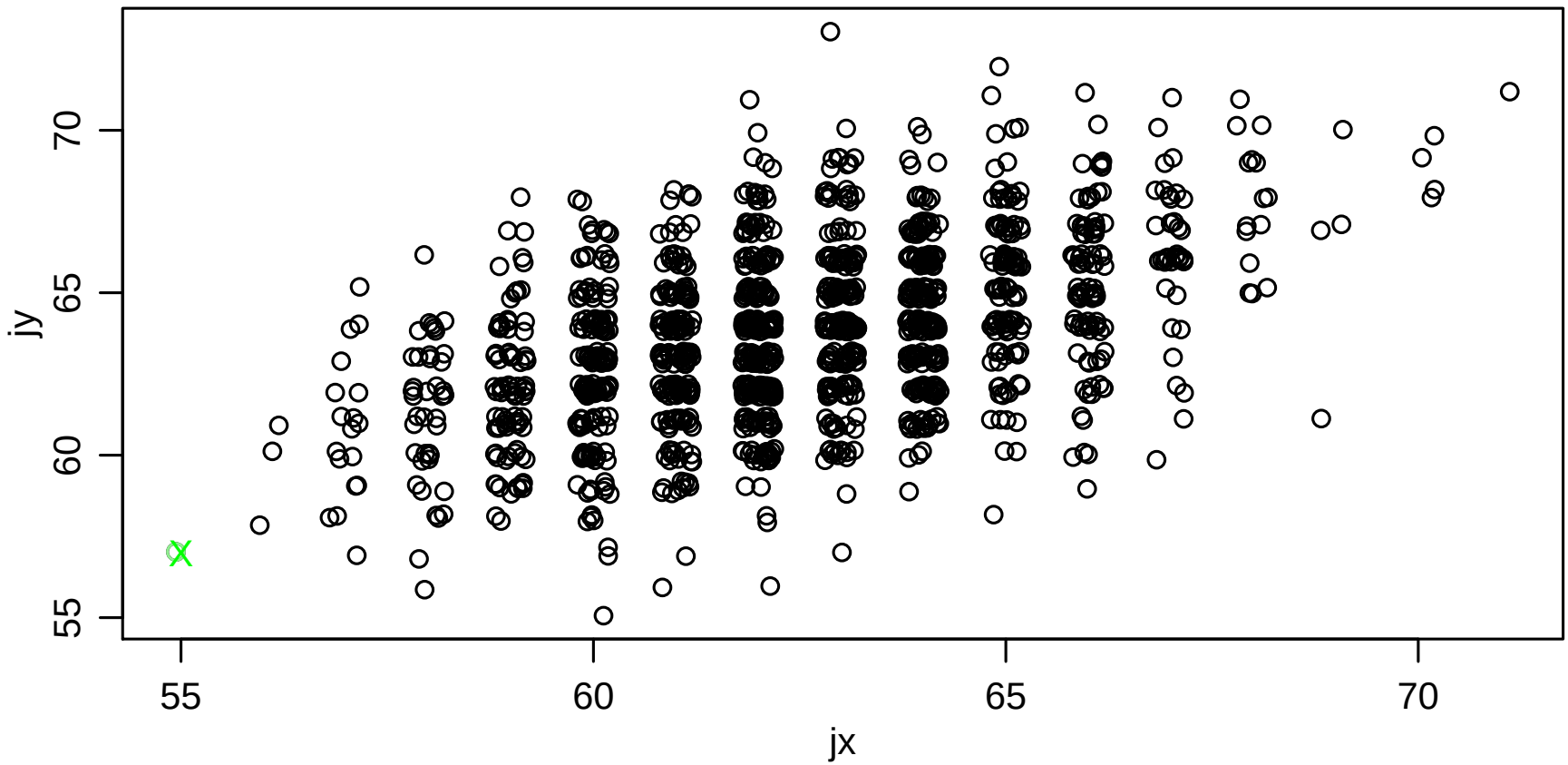
Linear model

Mean model: $\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x$, for any x

Unknown parameters: β_0, β_1 (two parameters)

Parameter estimates: $\hat{\beta}_0, \hat{\beta}_1$ (OLS values)

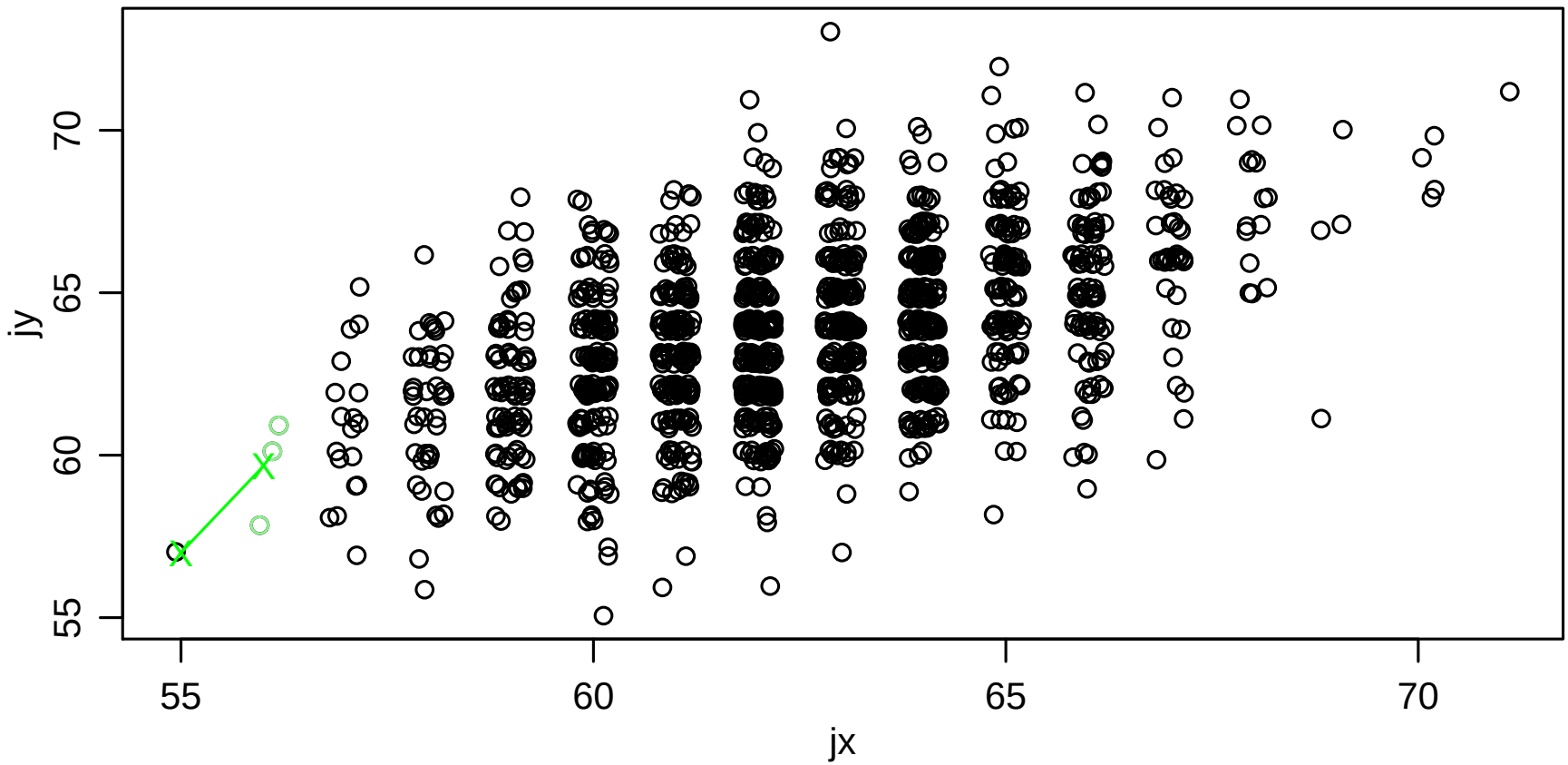
Estimating the unrestricted model




```
mean(y[x==55])
```

```
## [1] 57
```

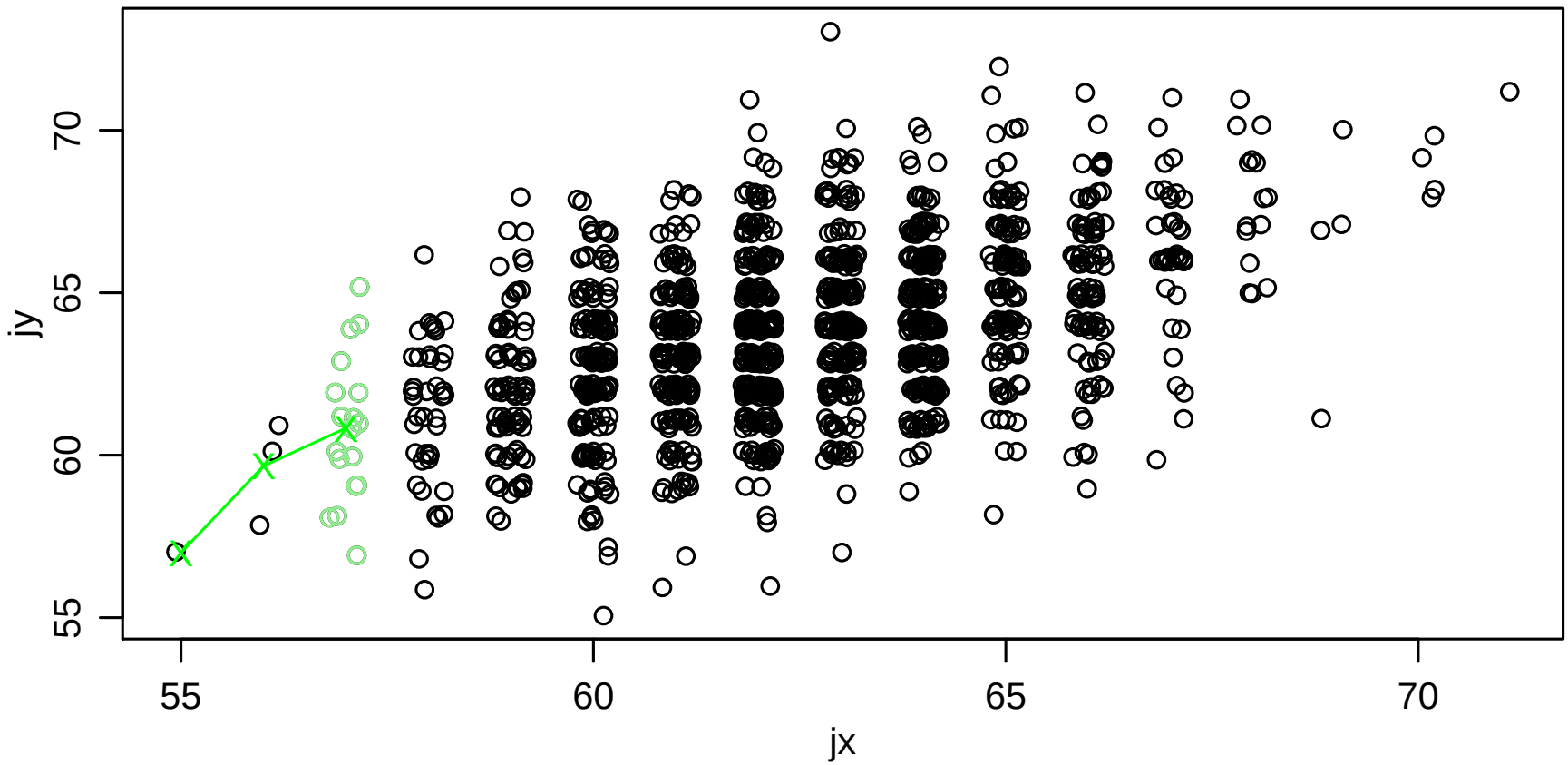
Estimating the unrestricted model




```
mean(y[x==56])
```

```
## [1] 59.66667
```

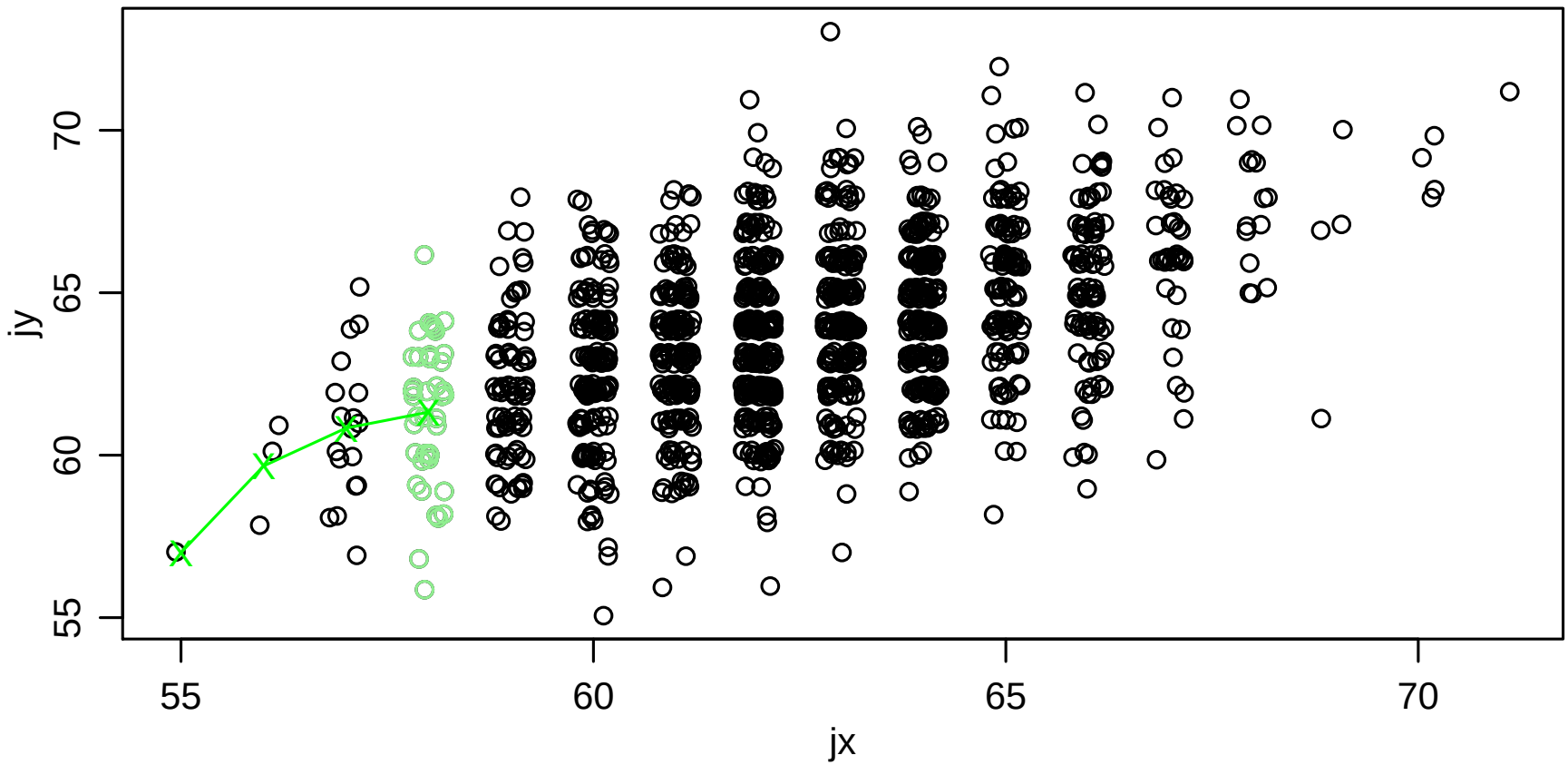
Estimating the unrestricted model




```
mean(y[x==57])
```

```
## [1] 60.83333
```

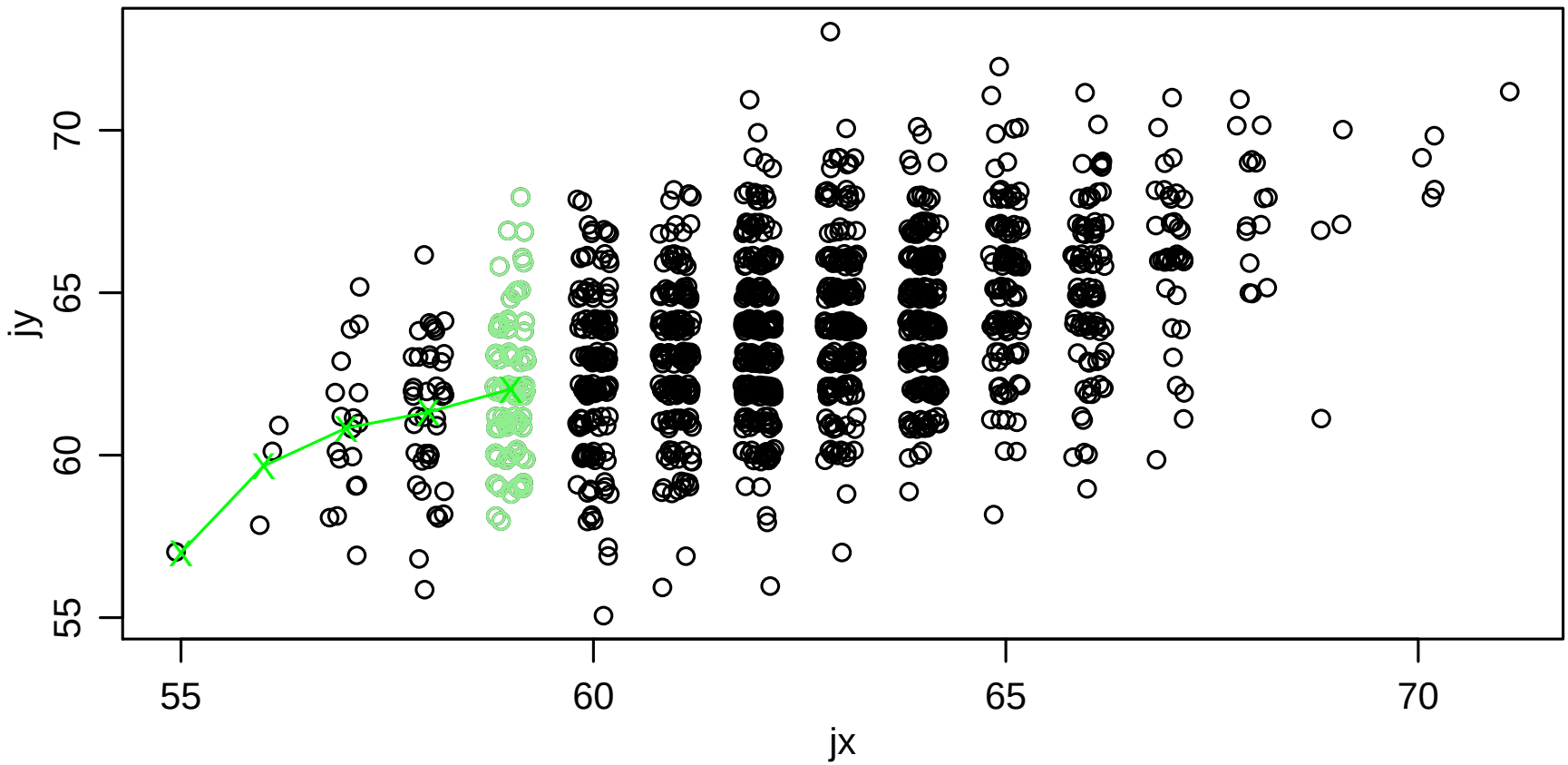
Estimating the unrestricted model




```
mean(y[x==58])
```

```
## [1] 61.325
```

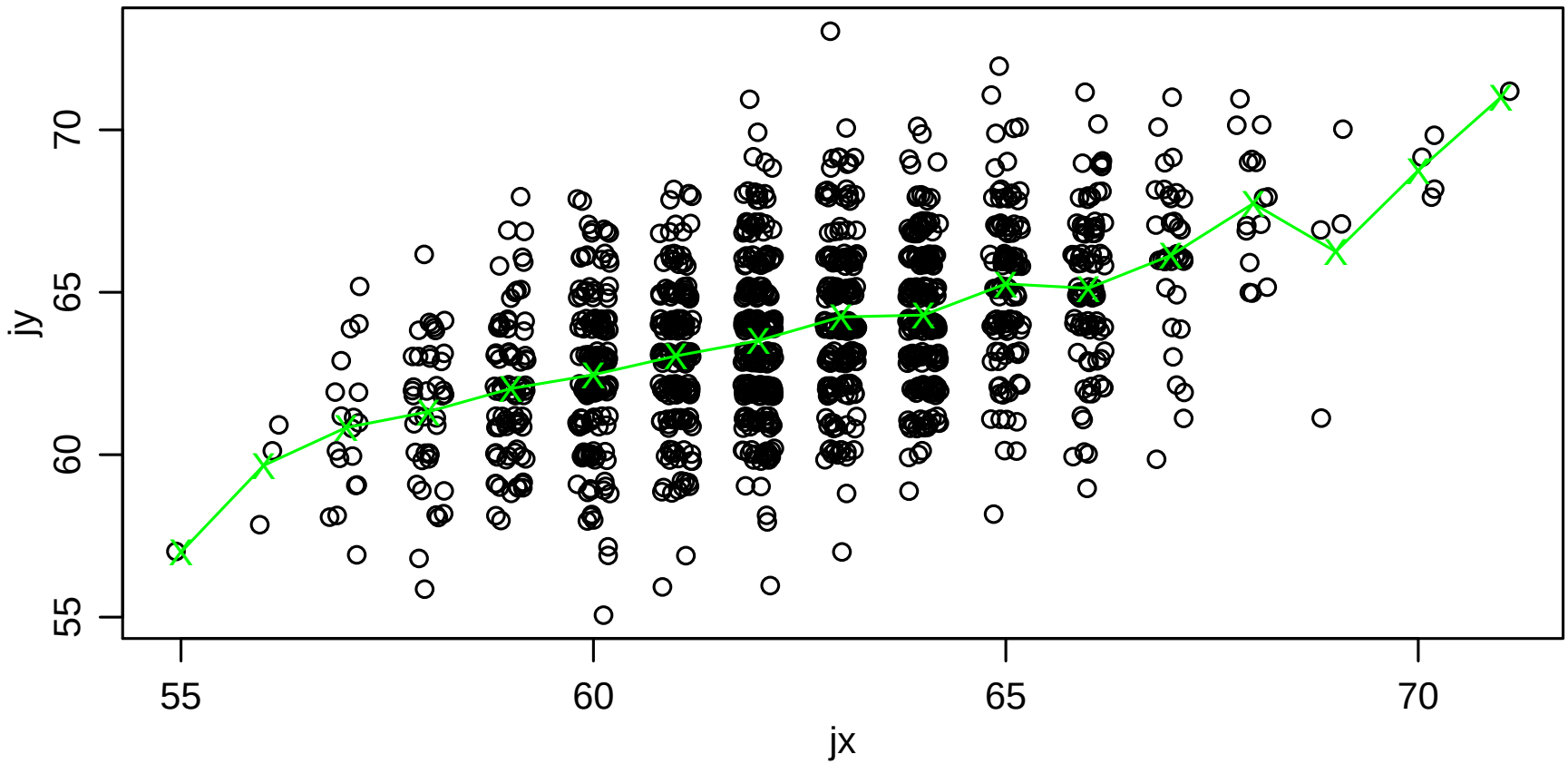
Estimating the unrestricted model




```
mean(y[x==59])
```

```
## [1] 62.03659
```

Estimating the unrestricted model

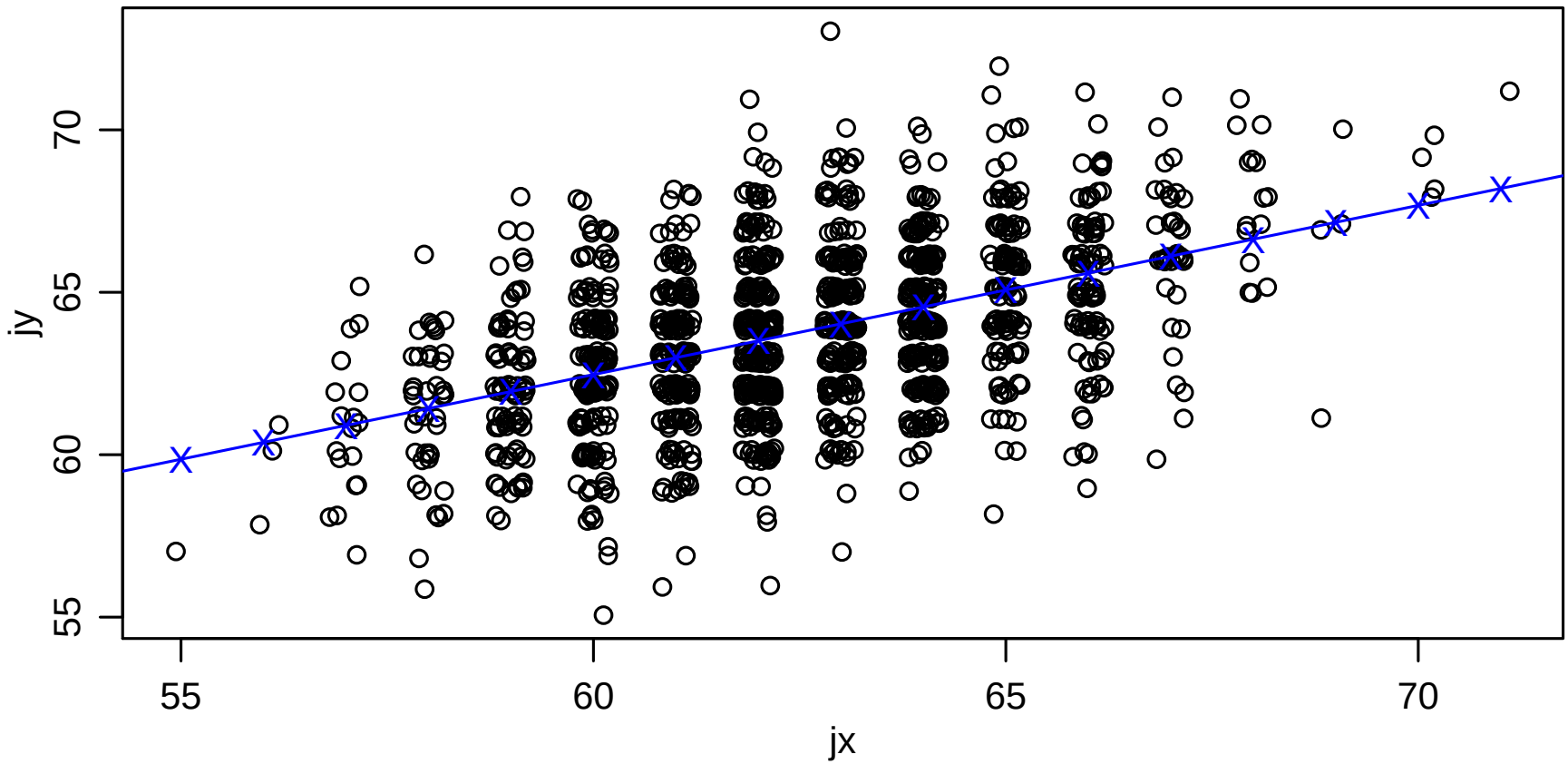



```
mux<-tapply(y,x,mean)

mux[1:5]

##       55       56       57       58       59
## 57.00000 59.66667 60.83333 61.32500 62.03659
```

Estimating the linear model




```
fit<-lm( y ~ x )

fit$coef

## (Intercept)           x
##  31.2082961   0.5209411
```

RSS

$$RSS(\hat{\mu}_{55}, \ldots, \hat{\mu}_{71}) = \sum_{i=1}^{n} (y_i - \hat{\mu}_{x_i})^2$$

$$RSS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_i])^2$$

Q: Which is bigger, $RSS(\hat{\mu}_{55}, \ldots, \hat{\mu}_{71})$ or $RSS(\hat{\beta}_0, \hat{\beta}_1)$?

RSS

$$RSS(\hat{\mu}_{55}, \ldots, \hat{\mu}_{71}) = \sum_{i=1}^{n} (y_i - \hat{\mu}_{x_i})^2$$

$$RSS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_i])^2$$

Q: Which is bigger, $RSS(\hat{\mu}_{55}, \ldots, \hat{\mu}_{71})$ or $RSS(\hat{\beta}_0, \hat{\beta}_1)$?

A: $RSS(\hat{\mu}_{55}, \ldots, \hat{\mu}_{71}) \leq RSS(\hat{\beta}_0, \hat{\beta}_1)$ (Homework!)

Residual sums of squares

```
mux<-tapply(y,x,mean)

ux<-sort(unique(x))

yfit_np<-mux[ match(x,ux) ]

mean( (y-yfit_np)^2 )

## [1] 5.266022
```

Residual sums of squares

```
bols<-lm(y~x)$coef

bols

## (Intercept)           x
##  31.2082961   0.5209411

yfit_ls<-bols[1]+bols[2]*x

mean( (y-yfit_ls)^2 )

## [1] 5.331938

mean(fit$res^2)

## [1] 5.331938
```

Prediction problem

Task: Given $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, predict y_{n+1} from x_{n+1}.

Prediction problem

Task: Given $\{(x_1, y_1), \ldots, (x_n, y_n)\}$, predict y_{n+1} from x_{n+1}.

Strategy:

1. Estimate $\mathrm{E}[Y|X = x_{n+1}]$ using $\{(x_1, y_1), \ldots, (x_n, y_n)\}$;
2. predict y_{n+1} with $\hat{y}_{n+1} = \widehat{\mathrm{E}[Y|X = x_{n+1}]}$.
   - For the unrestricted model, $\widehat{\mathrm{E}[Y|X = x_{n+1}]} = \hat{\mu}_{x_{n+1}}$
   - For the linear model, $\widehat{\mathrm{E}[Y|X = x_{n+1}]} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$.

Prediction error

We would like to know how good the predictions are on average:

$$\text{Prediction error} = \mathrm{E}[(Y_{n+1} - \hat{Y}_{n+1})^2]$$

Prediction error

We would like to know how good the predictions are on average:

$$\text{Prediction error} = \mathrm{E}[(Y_{n+1} - \hat{Y}_{n+1})^2]$$

However, we can't compute the expectation as the true population is unknown.

Prediction error

We would like to know how good the predictions are on average:

$$\text{Prediction error} = \mathrm{E}[(Y_{n+1} - \hat{Y}_{n+1})^2]$$

However, we can't compute the expectation as the true population is unknown.

Can we **estimate** the prediction error using the sample?

Cross validation

Procedure for each $i \in \{1, \ldots, n\}$

1. create $(\mathbf{x}_{[-i]}, \mathbf{y}_{[-i]})$, the dataset with (x_i, y_i) removed;
2. fit the mean model with $(\mathbf{x}_{[-i]}, \mathbf{y}_{[-i]})$;
3. predict y_i with $\hat{y}_i$.

Cross validation

Procedure for each $i \in \{1, \ldots, n\}$

1. create $(\mathbf{x}_{[-i]}, \mathbf{y}_{[-i]})$, the dataset with (x_i, y_i) removed;
2. fit the mean model with $(\mathbf{x}_{[-i]}, \mathbf{y}_{[-i]})$;
3. predict y_i with $\hat{y}_i$.

Estimate prediction error with $\widehat{PSS} = \frac{1}{n} \sum (y_i - \hat{y}_i)^2$

Cross validation

```
yprd_ls<-yprd_np<-NULL

for(i in 1:n)
{
  xmi<-x[-i]
  ymi<-y[-i]

  bmi<-lm(ymi~xmi)$coef ; yprd_ls<-c(yprd_ls, bmi[1] + bmi[2]*x[i] )

  yprd_np<-c(yprd_np,mean(ymi[xmi==x[i]]) )
}
```

Cross validation

```
yprd_ls<-yprd_np<-NULL

for(i in 1:n)
{
  xmi<-x[-i]
  ymi<-y[-i]

  bmi<-lm(ymi~xmi)$coef ; yprd_ls<-c(yprd_ls, bmi[1] + bmi[2]*x[i] )

  yprd_np<-c(yprd_np,mean(ymi[xmi==x[i]]) )
}
```

```
mean( (y-yprd_ls)^2 )

## [1] 5.347388

mean( (y-yprd_np)^2 )

## [1] NaN
```

Cross validation

```
yprd_ls<-yprd_np<-NULL

for(i in 1:n)
{
  xmi<-x[-i]
  ymi<-y[-i]

  bmi<-lm(ymi~xmi)$coef ; yprd_ls<-c(yprd_ls, bmi[1] + bmi[2]*x[i] )

  yprd_np<-c(yprd_np,mean(ymi[xmi==x[i]]) )
}
```

```
mean( (y-yprd_ls)^2 )

## [1] 5.347388

mean( (y-yprd_np)^2 )

## [1] NaN
```

```
mean( (y-yprd_ls)[!is.na(yprd_np)]^2, na.rm=TRUE )

## [1] 5.34328

mean( (y-yprd_np)[!is.na(yprd_np)]^2, na.rm=TRUE )

## [1] 5.394983
```

Questions and comments

Questions:

1. Do you think that the true $\mathrm{E}[Y|X = x]$ is exactly linear in x?
2. Do we have enough data to reliably estimate a model with no restrictions?

Questions and comments

Questions:

1. Do you think that the true $\mathrm{E}[Y|X = x]$ is exactly linear in x?
2. Do we have enough data to reliably estimate a model with no restrictions?

Comments:

1. The appropriateness of the linear model depends on the truth and the amount of data.
2. Other possible models are polynomial regression, splines, smoothers.