S&DS 355 / 365 / 565

(Schrodinger's Intro) Machine Learning and Data Mining

Linear Regression

Tuesday, September 3rd

Yale

365 + 355

For the next week, we will be merging both 365 and 355 classes.

365 vs 355

- 355 does not count towards the S&DS Major
- 365 = more math, harder topics, 355 = more programming

For today

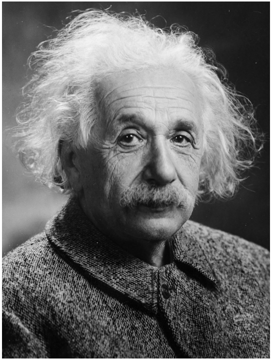
- Linear regression
  - Estimation
  - Measures of fit
  - Inference
  - Several predictors
- Redux of part of last lecture

Why start here?

- Linear regression is foundation for more sophisticated topics:
  - Regularization: Ridge regression and lasso
  - Kernel methods: Support vector machines
  - Smoothing: Splines, generalized additive models, etc.

Why start here?

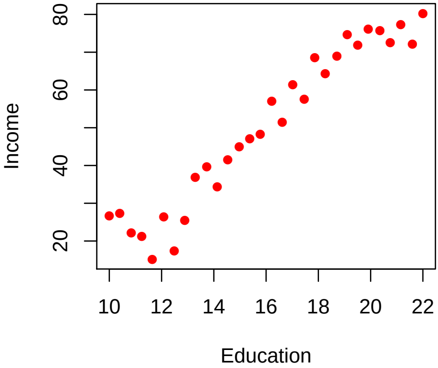
- Linear regression is foundation for more sophisticated topics:
  - Regularization: Ridge regression and lasso
  - Kernel methods: Support vector machines
  - Smoothing: Splines, generalized additive models, etc.
- Many advanced machine learning methods are generalizations or extensions of linear regression



Everything should be made as simple as possible, but no simpler.

https://en.wikiquote.org/wiki/Albert_Einstein

Simulated income dataset

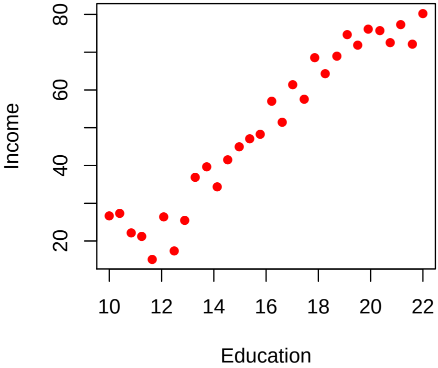


Goal: Predict income(Y) using education (X).

$$Y = f(X) + \epsilon$$

$(x_1, y_1), (x_2, y_2), \ldots, (x_{30}, y_{30})$

Simulated income dataset



Goal: Predict income(Y) using education (X).

$$Y = f(X) + \epsilon$$

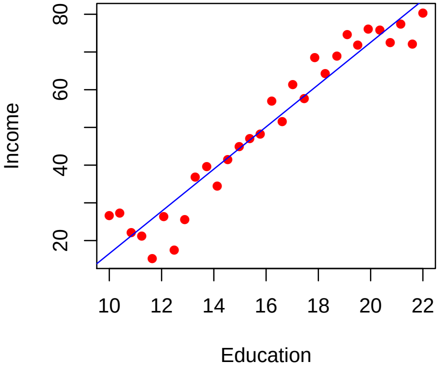
Linear model:

$$f(X) = \beta_0 + \beta_1 X$$

$$\epsilon \sim N(0, \sigma^2)$$

$(x_1, y_1), (x_2, y_2), \ldots, (x_{30}, y_{30})$

Simulated income dataset



Goal: Predict income(Y) using education (X).

$$Y = f(X) + \epsilon$$

Linear model:

$$f(X) = \beta_0 + \beta_1 X$$

$$\epsilon \sim N(0, \sigma^2)$$

Find coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$ s.t. $\widehat{Y} = \widehat{f}(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X$ is reasonably close to Y.

Estimating the coefficients

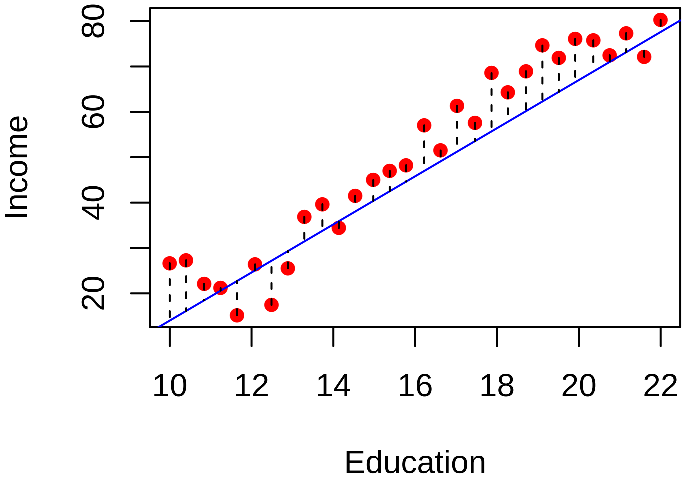
For any $\widehat{\beta}_0, \widehat{\beta}_1$, we predict $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$. We call these **fitted values.**

Estimating the coefficients

For any $\widehat{\beta}_0, \widehat{\beta}_1$, we predict $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$. We call these **fitted values.**

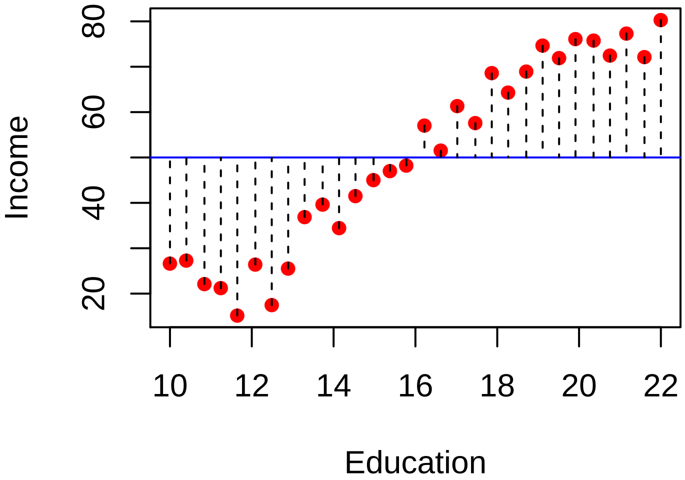
The **residual** $e_i = y_i - \widehat{y}_i$ is difference between the i-th observed value and its fitted value.

Some candidate lines (and residuals)



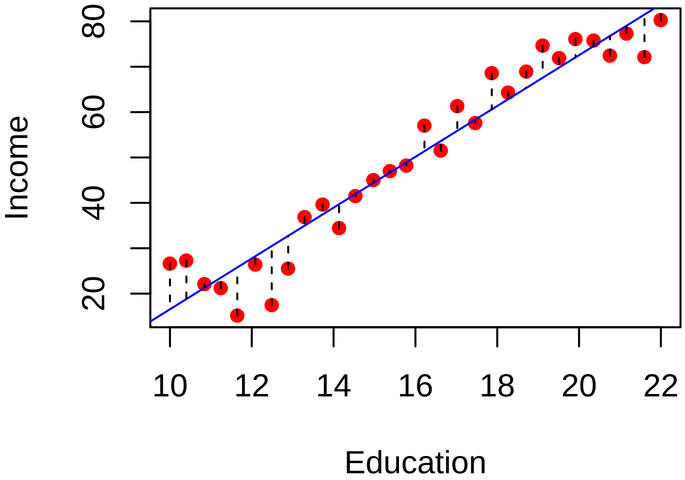
$$\widehat{\beta}_0 = -39, \widehat{\beta}_1 = 5.3$$

Some candidate lines (and residuals)



$\widehat{\beta}_0 = 50, \widehat{\beta}_1 = 0$

Some candidate lines (and residuals)



$\widehat{\beta}_0 = -39.4, \widehat{\beta}_1 = 5.6$

Estimating the coefficients

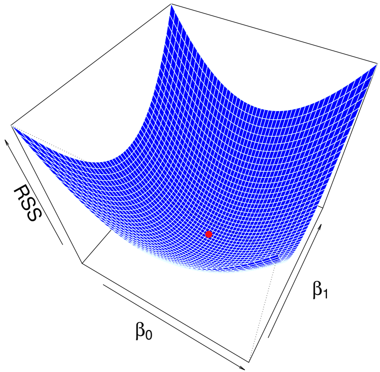
The **least squares** approach selects coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that minimize the **residual sum of squares** (RSS):

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2.$$

Estimating the coefficients

The **least squares** approach selects coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that minimize the **residual sum of squares** (RSS):

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = (y_1 - \beta_0 - \beta_1 x_1)^2 + \cdots + (y_n - \beta_0 - \beta_1 x_n)^2.$$



Estimating the coefficients

The **least squares** approach selects coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that minimize the **residual sum of squares** (RSS):

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = (y_1 - \beta_0 - \beta_1 x_1)^2 + \cdots + (y_n - \beta_0 - \beta_1 x_n)^2.$$

How do we find the minimum?

- $RSS(\beta_0, \beta_1)$ is convex.
- Take partial derivatives and set to 0:

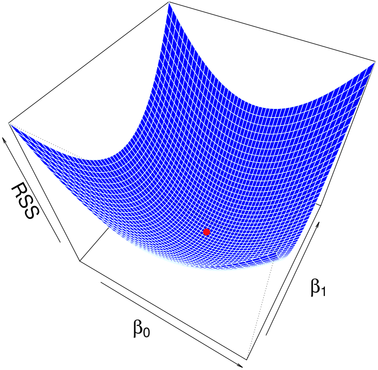
$$\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial RSS(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} -2x_i(y_i - \beta_0 - \beta_1 x_i) = 0$$

Estimating the coefficients

The **least squares** approach selects coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that minimize the **residual sum of squares** (RSS):

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = (y_1 - \beta_0 - \beta_1 x_1)^2 + \cdots + (y_n - \beta_0 - \beta_1 x_n)^2.$$



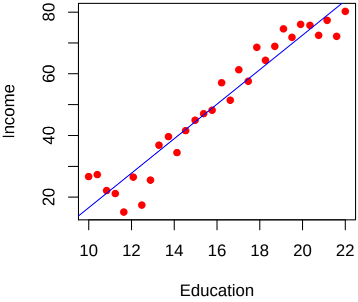
Minimum RSS is achieved at:

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x},$$

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

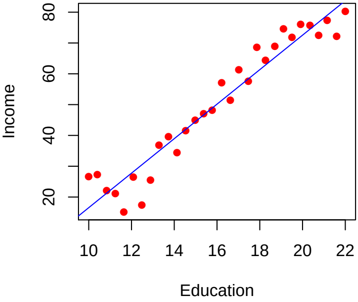
where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Simulated income dataset



$$\widehat{\beta}_0 = -39.45 \qquad \widehat{\beta}_1 = 5.60$$

Simulated income dataset



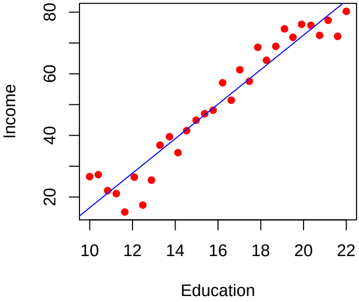
$$\widehat{\beta}_0 = -39.45 \qquad \widehat{\beta}_1 = 5.60$$

$$\widehat{y} = -39.45 + 5.60x$$

Interpretation:

- A one-year increase in education is associated with an increase in average income of 5.6 units.

Simulated income dataset



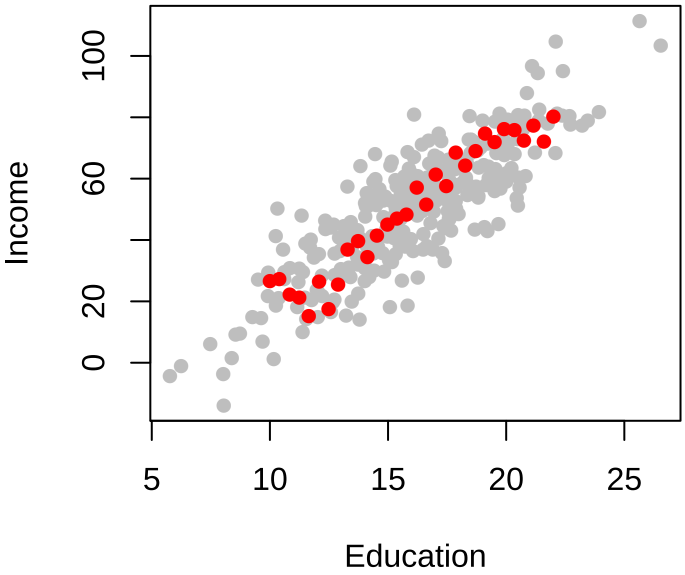
$$\widehat{\beta}_0 = -39.45 \qquad \widehat{\beta}_1 = 5.60$$

$$\widehat{Income} = -39.45 + 5.60 \cdot Education$$

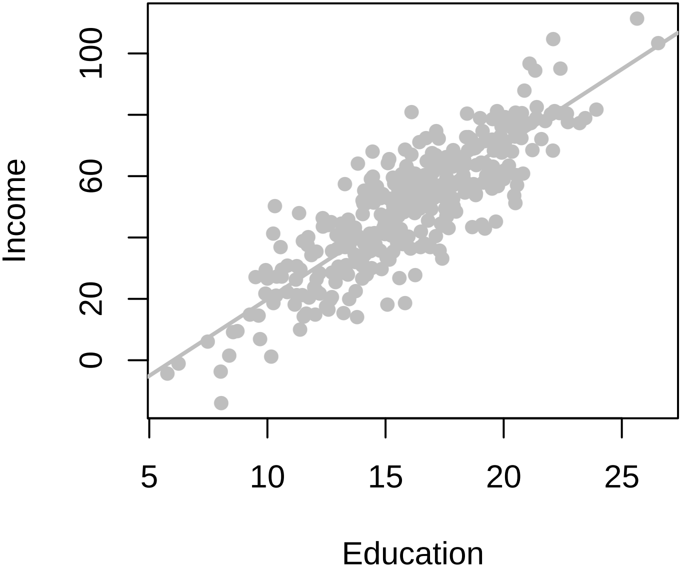
Interpretation:

- A one-year increase in education is associated with an increase in average income of 5.6 units.

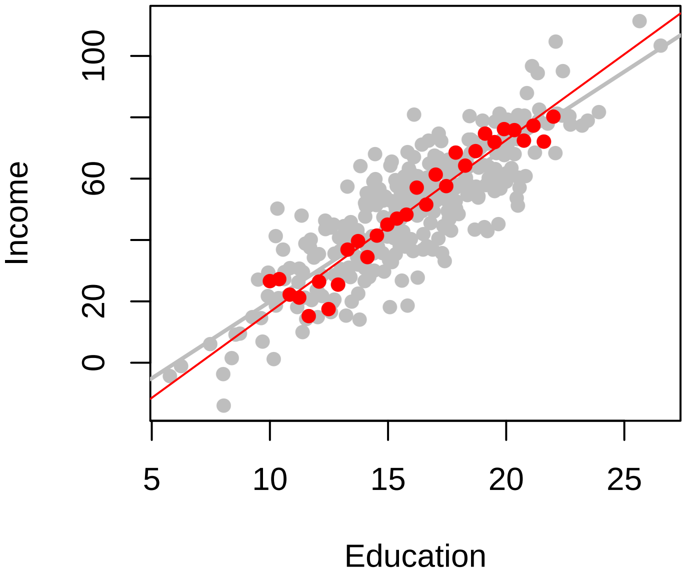
Population vs. sample



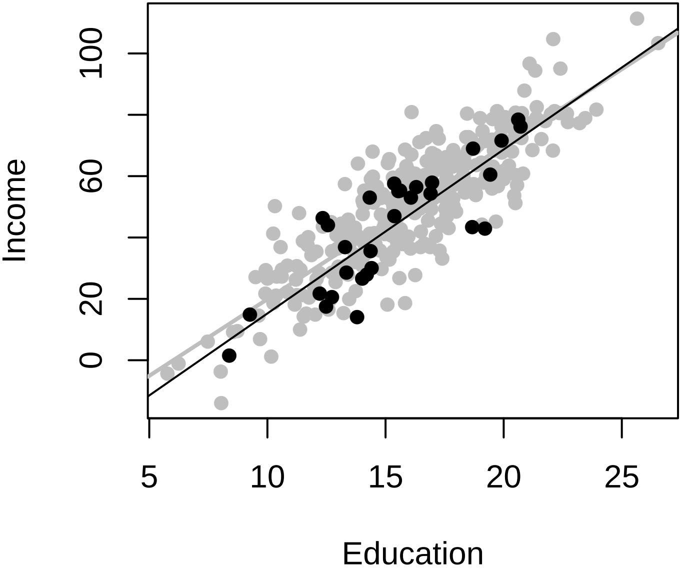
Population vs. sample



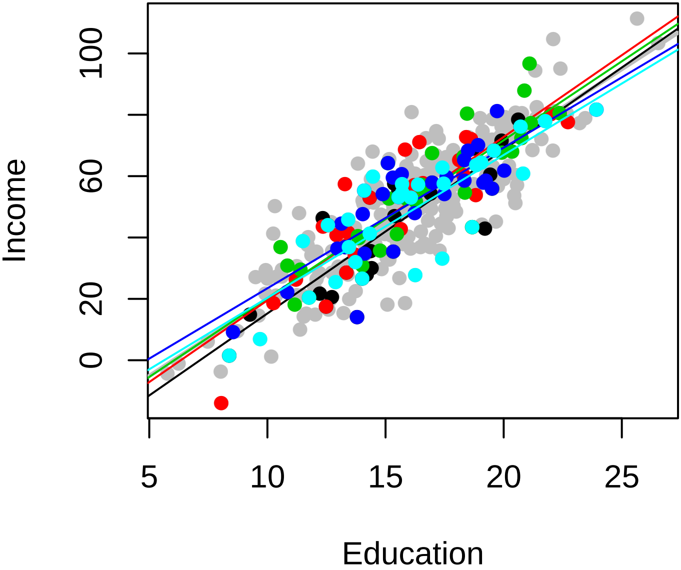
Population vs. sample



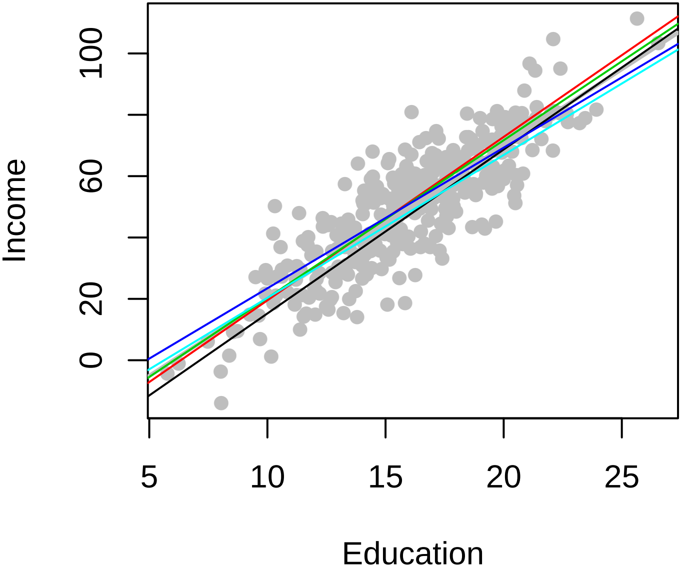
Different samples



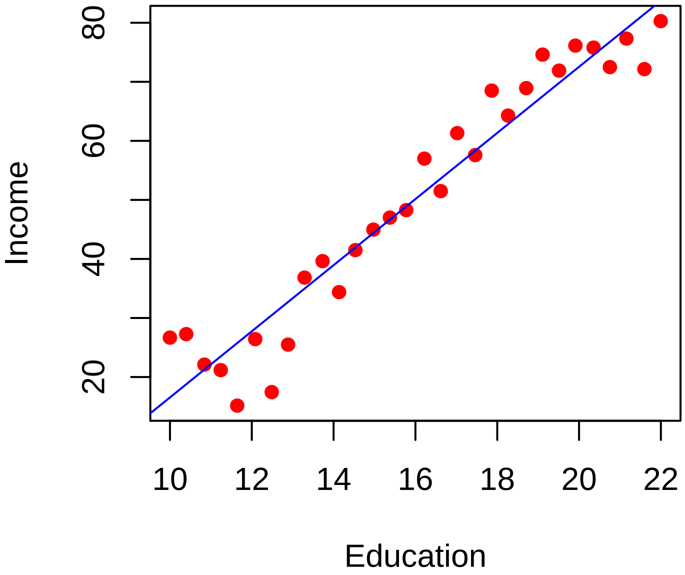
Different samples



Different samples



How to determine variability?



Inference for linear regression

Standard errors of the coefficients describe how the coefficients vary under repeated sampling:

$$SE(\widehat{\beta}_0) = \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$$

$$SE(\widehat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where $\sigma^2 = Var(\epsilon)$.

Inference for linear regression

Standard errors of the coefficients describe how the coefficients vary under repeated sampling:

$$SE(\widehat{\beta}_0) = \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$$

$$SE(\widehat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

where $\sigma^2 = Var(\epsilon)$. A 95% confidence interval for β_i is approximately:

$$\widehat{\beta}_i \pm 2 \cdot SE(\widehat{\beta}_i)$$

Inference for linear regression

```
##
## Call:
## lm(formula = Income ~ Education, data = x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.046  -2.293   0.472   3.288  10.110
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -39.4463     4.7248  -8.349  4.4e-09 ***
## Education     5.5995     0.2882  19.431  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.653 on 28 degrees of freedom
## Multiple R-squared:  0.931,Adjusted R-squared:  0.9285
## F-statistic: 377.6 on 1 and 28 DF,  p-value: < 2.2e-16
```

Inference for linear regression

A 95% confidence interval for β_1 is approximately:

$$5.6 \pm 2 \times 0.288 = (5.0, 6.2)$$

Inference for linear regression

A 95% confidence interval for β_1 is approximately:

$$5.6 \pm 2 \times 0.288 = (5.0, 6.2)$$

For a one-year increase in an individual's education, the model predicts that the individual's income will rise by between 5.0 and 6.2 units.

Inference for linear regression

A 95% confidence interval for β_1 is approximately:

$$5.6 \pm 2 \times 0.288 = (5.0, 6.2)$$

For a one-year increase in an individual's education, the model predicts that the individual's income will rise by between 5.0 and 6.2 units. Note: This statement is about the population slope β_1!

Sums of squares and R^2

Partitioning the sums of squares:

$$\underbrace{\sum (y_i - \bar{y})^2}_{\text{total sum of squares}(TSS)} = \underbrace{\sum (\widehat{y}_i - \bar{y})^2}_{\text{explained sum of squares}(ESS)} + \underbrace{\sum (y_i - \widehat{y}_i)^2}_{\text{residual sum of squares}(RSS)}$$

(need some algebra to show this)

Sums of squares and R^2

Partitioning the sums of squares:

$$\underbrace{\sum (y_i - \bar{y})^2}_{\text{total sum of squares}(TSS)} = \underbrace{\sum (\widehat{y}_i - \bar{y})^2}_{\text{explained sum of squares}(ESS)} + \underbrace{\sum (y_i - \widehat{y}_i)^2}_{\text{residual sum of squares}(RSS)}$$

(need some algebra to show this)

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

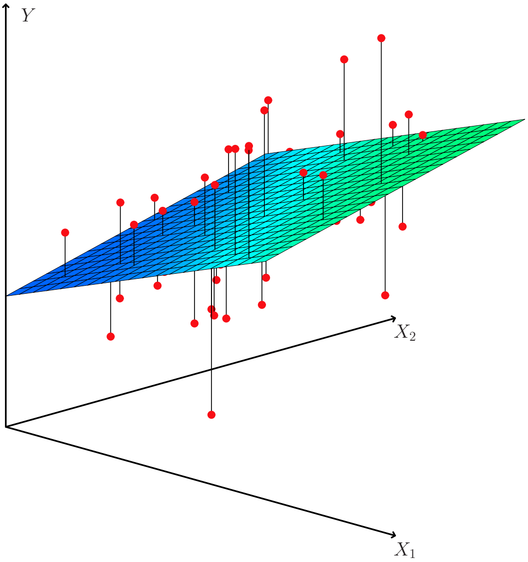
We can interpret R^2 (**multiple R-squared**) as the proportion of variability in y explained by the model.

- Between 0 and 1
- Doesn't depend on the scale of Y.

Inference for linear regression

```
##
## Call:
## lm(formula = Income ~ Education, data = x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.046  -2.293   0.472   3.288  10.110
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -39.4463     4.7248  -8.349  4.4e-09 ***
## Education     5.5995     0.2882  19.431  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.653 on 28 degrees of freedom
## Multiple R-squared:  0.931,Adjusted R-squared:  0.9285
## F-statistic: 377.6 on 1 and 28 DF,  p-value: < 2.2e-16
```

Multiple linear regression



General form

With p predictors $x_1, \ldots, x_p$,

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$.

In matrix notation,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & \ddots & & x_{2,p} \\ \vdots & & \ddots & \vdots & \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

General form

With p predictors $x_1, \ldots, x_p$,

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$.

In matrix notation,

$$y = X\beta + \epsilon$$

(where the intercept β_0 corresponds to a column of all 1s)

Residual sum of squares

Recall that

$$\widehat{\beta} = \arg\min_{\beta} RSS(\beta).$$

Residual sum of squares

Recall that

$$\widehat{\beta} = \arg\min_{\beta} RSS(\beta).$$

$$\begin{aligned}
RSS(\beta) &= \|y - X\beta\|_2^2 \\
&= (y - X\beta)^T (y - X\beta) \\
&= (y^T - \beta^T X^T)(y - X\beta) \\
&= y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta \\
&= y^T y - 2y^T X\beta + \beta^T X^T X\beta
\end{aligned}$$

Residual sum of squares

Recall that

$$\widehat{\beta} = \arg\min_{\beta} RSS(\beta).$$

$$\begin{aligned}
RSS(\beta) &= \|y - X\beta\|_2^2 \\
&= (y - X\beta)^T (y - X\beta) \\
&= (y^T - \beta^T X^T)(y - X\beta) \\
&= y^T y - y^T X\beta - \beta^T X^T y + \beta^T X^T X\beta \\
&= y^T y - 2y^T X\beta + \beta^T X^T X\beta
\end{aligned}$$

Hence,

$$\begin{aligned}
\widehat{\beta} &= \arg\min_{\beta} \left[-2y^T X\beta + \beta^T X^T X\beta\right] \\
&= \arg\min_{\beta} \left[-2\tfrac{1}{n} y^T X\beta + \beta^T \tfrac{1}{n} X^T X\beta\right].
\end{aligned}$$

Estimating β

Compute derivatives of $RSS(\beta)$ with respect to β_i and set equal to 0.

Estimating β

Compute derivatives of $RSS(\beta)$ with respect to β_i and set equal to 0.

The β that minimizes $RSS(\beta)$ satisfies the **normal equations**:

$$X^T X\beta = X^T y.$$

Estimating β

Compute derivatives of $RSS(\beta)$ with respect to β_i and set equal to 0.

The β that minimizes $RSS(\beta)$ satisfies the **normal equations**:

$$X^T X\beta = X^T y.$$

If the matrix $X^T X$ is invertible, solve to get

$$\widehat{\beta} = (X^T X)^{-1} X^T y.$$

Interpretation

The coefficients are just the correlations between the variables X_j and the data Y—*after* the variables are "whitened" to become uncorrelated.

For the algebraically inclined

Let $\widetilde{X} = X\widehat{\Sigma}^{-1/2}$, where $\widehat{\Sigma} = \frac{1}{n}X^T X$ is the sample covariance.

For the algebraically inclined

Let $\widetilde{X} = X\widehat{\Sigma}^{-1/2}$, where $\widehat{\Sigma} = \frac{1}{n}X^T X$ is the sample covariance.

Then $\widetilde{X}$ is "whitened" (uncorrelated):

$$\begin{aligned}\frac{1}{n}\widetilde{X}^T\widetilde{X} &= \widehat{\Sigma}^{-1/2}\left(\frac{1}{n}X^T X\right)\widehat{\Sigma}^{-1/2} \\ &= \widehat{\Sigma}^{-1/2}\widehat{\Sigma}\widehat{\Sigma}^{-1/2} \\ &= \widehat{\Sigma}^{1/2}\widehat{\Sigma}^{-1/2} \\ &= I\end{aligned}$$

For the algebraically inclined

Let $\widetilde{X} = X\widehat{\Sigma}^{-1/2}$, where $\widehat{\Sigma} = \frac{1}{n}X^T X$ is the sample covariance.

Then $\widetilde{X}$ is "whitened" (uncorrelated):

$$\begin{aligned}
\frac{1}{n}\widetilde{X}^T\widetilde{X} &= \widehat{\Sigma}^{-1/2}\left(\frac{1}{n}X^T X\right)\widehat{\Sigma}^{-1/2} \\
&= \widehat{\Sigma}^{-1/2}\widehat{\Sigma}\widehat{\Sigma}^{-1/2} \\
&= \widehat{\Sigma}^{1/2}\widehat{\Sigma}^{-1/2} \\
&= I
\end{aligned}$$

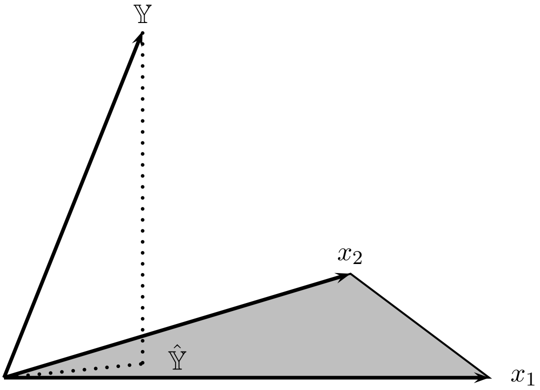
So the regression coefficients are

$$\widetilde{\beta} = \frac{1}{n}\widetilde{X}^T Y$$

which is (proportional to) correlation of $\widetilde{X}$ with Y. Transforming back gives $\widehat{\beta}$.

For the geometrically inclined

The **predicted values** (aka **fitted values**) $\widehat{Y} = X\widehat{\beta}$ are the orthogonal projection of the data $Y \in \mathbb{R}^n$ onto the column space of X (the span of columns $X_1, X_2, \ldots, X_p \in \mathbb{R}^n$)



Potential issues

$$\widehat{\beta} = (X^T X)^{-1} X^T y.$$

- $X^T X$ may not be invertible (if $n < p$)
- Inverting $X^T X$ can be computationally intensive, $O(p^3)$

Multiple linear regression

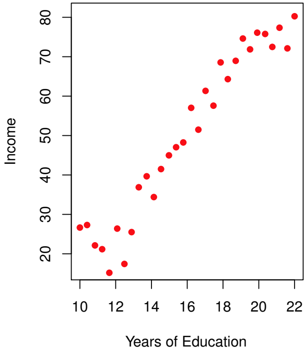
```
##
## Call:
## lm(formula = Income ~ ., data = x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.113 -5.718 -1.095  3.134 17.235
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -50.08564    5.99878  -8.349 5.85e-09 ***
## Education     5.89556    0.35703  16.513 1.23e-15 ***
## Seniority     0.17286    0.02442   7.079 1.30e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.187 on 27 degrees of freedom
## Multiple R-squared:  0.9341,Adjusted R-squared:  0.9292
## F-statistic: 191.4 on 2 and 27 DF,  p-value: < 2.2e-16
```

Comparing methods

Let's now go over some of the discussion from last lecture

Regression example

Back to regression with $p = 1$:



$$Y = f(X) + \epsilon$$

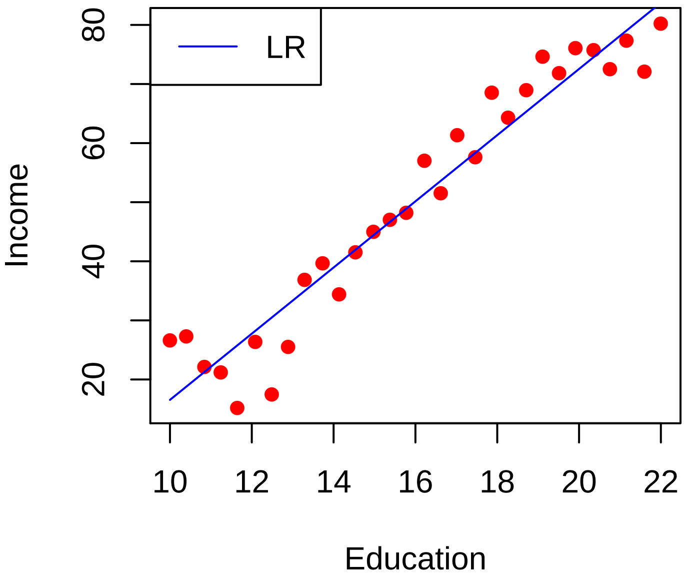
Modeling:

Use a procedure to get $\widehat{f}$. Derive estimates $\widehat{Y} = \widehat{f}(X)$.

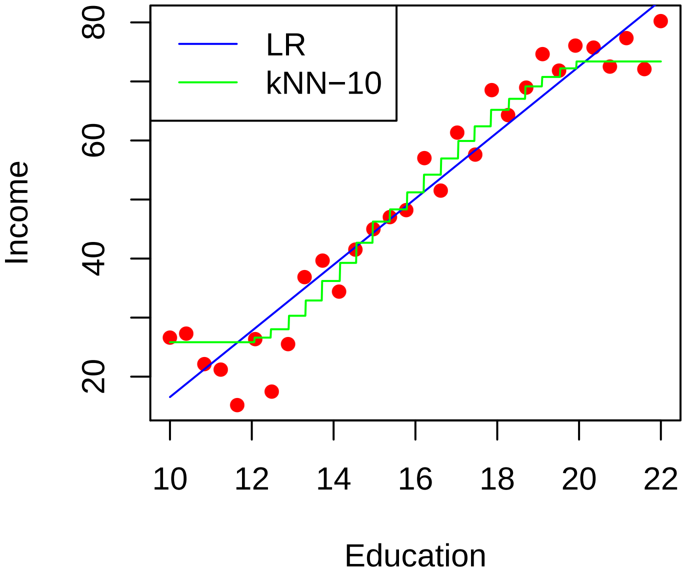
Possible regression approaches

- linear regression
  - Fitting a straight line through the data.
- *k*-nearest neighbors regression
  - Average together the y_i for x_i close to x

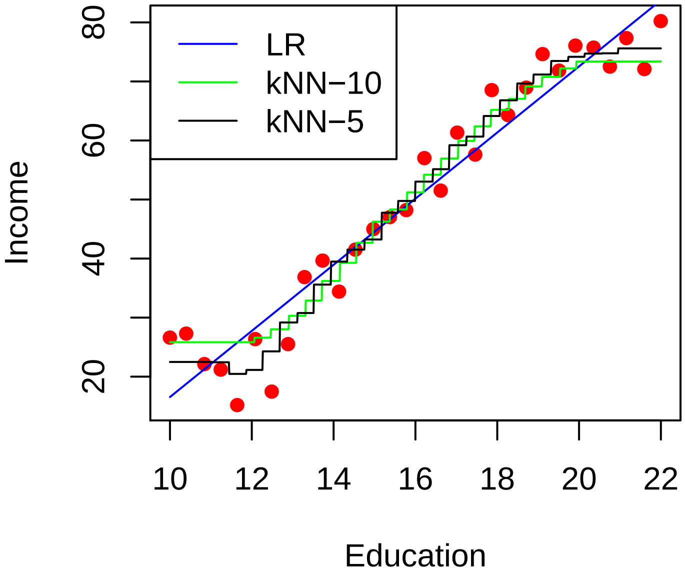
Possible regression approaches



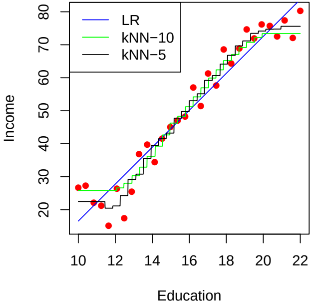
Possible regression approaches



Possible regression approaches



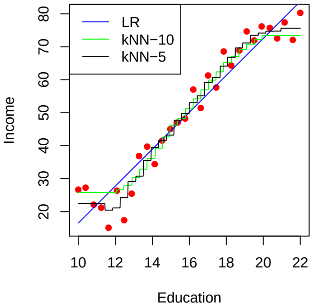
Possible regression approaches



Measuring performance via **Mean Squared Error**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2$$

Possible regression approaches



Measuring performance via **Mean Squared Error**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}(x_i))^2$$

MSEs for three methods:

Linear Regression	29.829
k-Nearest Neighbors (k=10)	23.519
k-Nearest Neighbors (k=5)	16.21

A k-nearest neighbors model with $k = 5$ achieves lowest error. Is it the best?

Training MSE vs. Test MSE

MSE in the previous table, **training MSE**, was computed based on data used in fitting the model.

We are more interested in **test MSE** computed on *unseen data*.

Training MSE vs. Test MSE

MSE in the previous table, **training MSE**, was computed based on data used in fitting the model.

We are more interested in **test MSE** computed on *unseen data*.What if we don't have other data?

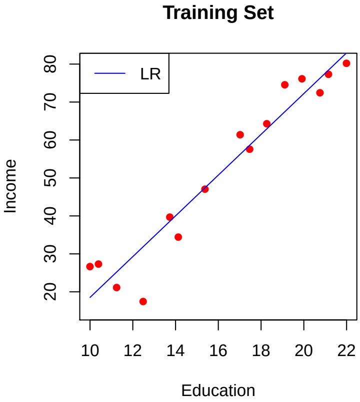
Training MSE vs. Test MSE

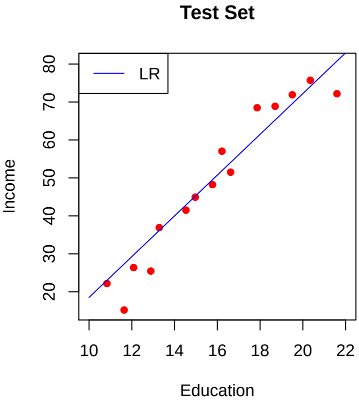
MSE in the previous table, **training MSE**, was computed based on data used in fitting the model.

We are more interested in **test MSE** computed on *unseen data*.What if we don't have other data?

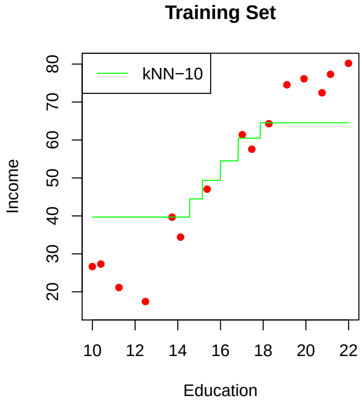
We can randomly split our data into a test set and a training set.

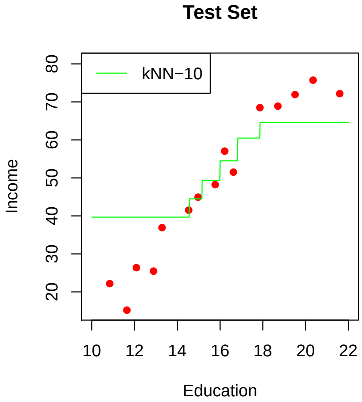
Regression approaches revisited



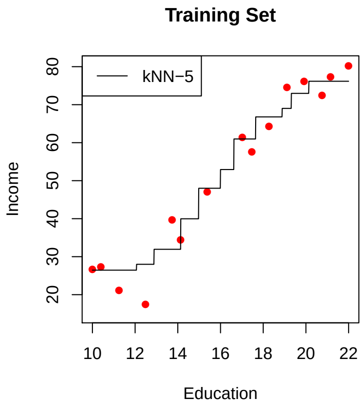


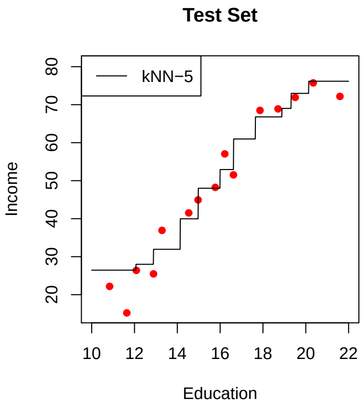
Regression approaches revisited



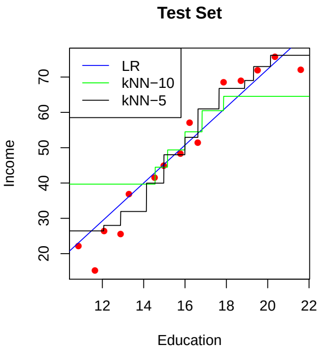


Regression approaches revisited





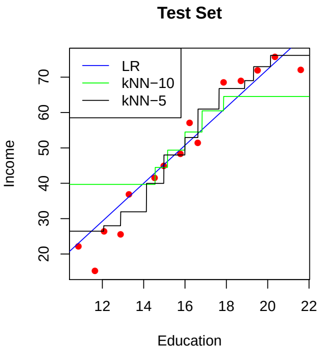
Regression approaches revisited



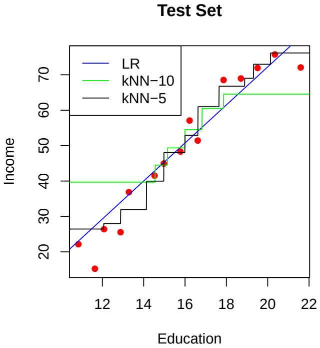
Regression approaches revisited

Compute MSE on the test set:

$$MSE = \frac{1}{n}\sum(y_i - \widehat{f}(x_i))^2$$



Regression approaches revisited

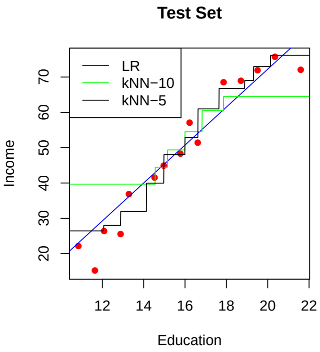


Compute MSE on the test set:

$$MSE = \frac{1}{n}\sum(y_i - \widehat{f}(x_i))^2$$

Linear Regression	28.502
k-Nearest Neighbors (k=10)	105.783
k-Nearest Neighbors (k=5)	24.898

Regression approaches revisited



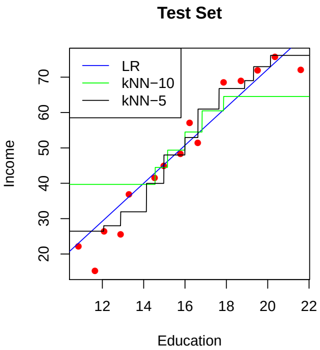
Compute MSE on the test set:

$$MSE = \frac{1}{n}\sum(y_i - \widehat{f}(x_i))^2$$

Linear Regression	28.502
k-Nearest Neighbors (k=10)	105.783
k-Nearest Neighbors (k=5)	24.898

So it appears that linear regression wins.

Regression approaches revisited



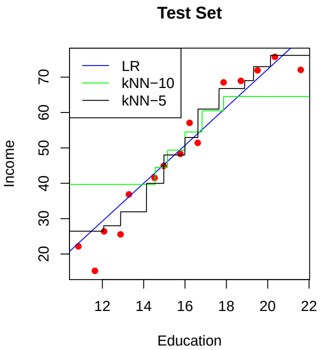
Compute MSE on the test set:

$$MSE = \frac{1}{n} \sum (y_i - \widehat{f}(x_i))^2$$

Linear Regression	28.502
k-Nearest Neighbors (k=10)	105.783
k-Nearest Neighbors (k=5)	24.898

So it appears that linear regression wins. Does it?

Regression approaches revisited



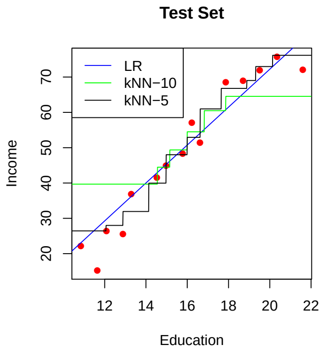
Compute MSE on the test set:

$$MSE = \frac{1}{n} \sum (y_i - \widehat{f}(x_i))^2$$

Linear Regression	28.502
k-Nearest Neighbors (k=10)	105.783
k-Nearest Neighbors (k=5)	24.898

So it appears that linear regression wins. Does it?

Regression approaches revisited



Compute MSE on the test set:

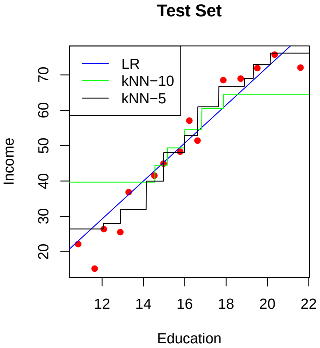
$$MSE = \frac{1}{n} \sum (y_i - \widehat{f}(x_i))^2$$

Linear Regression	28.502
k-Nearest Neighbors (k=10)	105.783
k-Nearest Neighbors (k=5)	24.898

So it appears that linear regression wins. Does it?

With different random splits of test vs. training, we could have gotten different results.

Regression approaches revisited



Compute MSE on the test set:

$$MSE = \frac{1}{n} \sum (y_i - \widehat{f}(x_i))^2$$

Linear Regression	28.502
k-Nearest Neighbors (k=10)	105.783
k-Nearest Neighbors (k=5)	24.898

So it appears that linear regression wins. Does it?

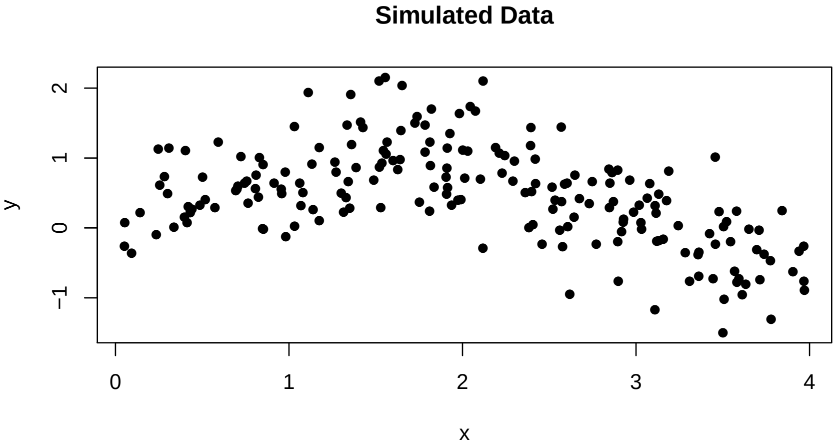
With different random splits of test vs. training, we could have gotten different results. We'll talk about ways around this later.

Overfitting

A method is **overfitting** the data when it has a small training MSE but a large test MSE.

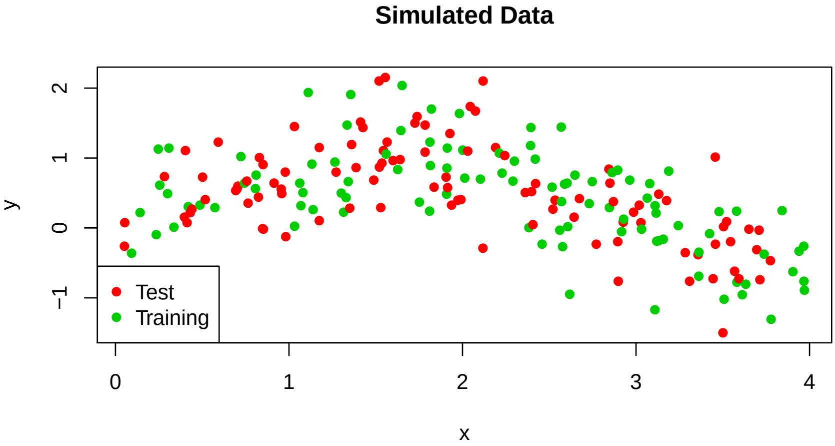
Overfitting

A method is **overfitting** the data when it has a small training MSE but a large test MSE. Using a bigger dataset:



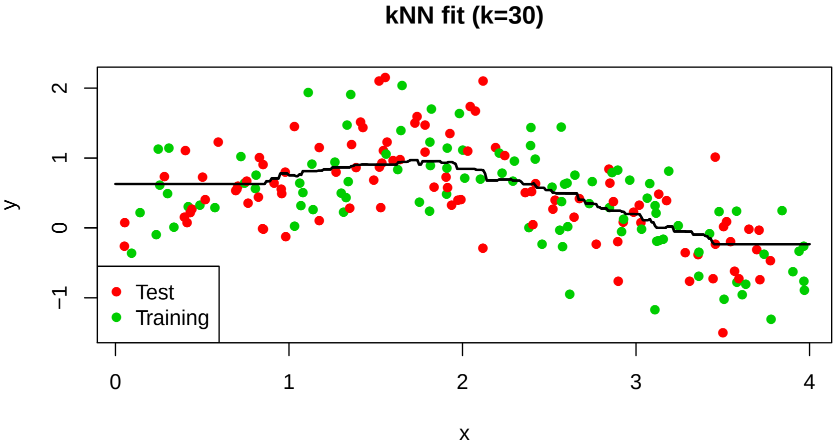
Overfitting

A method is **overfitting** the data when it has a small training MSE but a large test MSE. Using a bigger dataset:



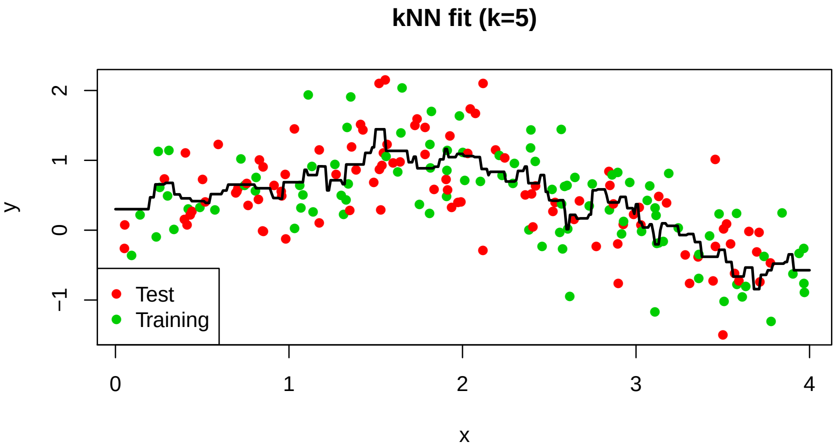
Overfitting

A method is **overfitting** the data when it has a small training MSE but a large test MSE. Using a bigger dataset:

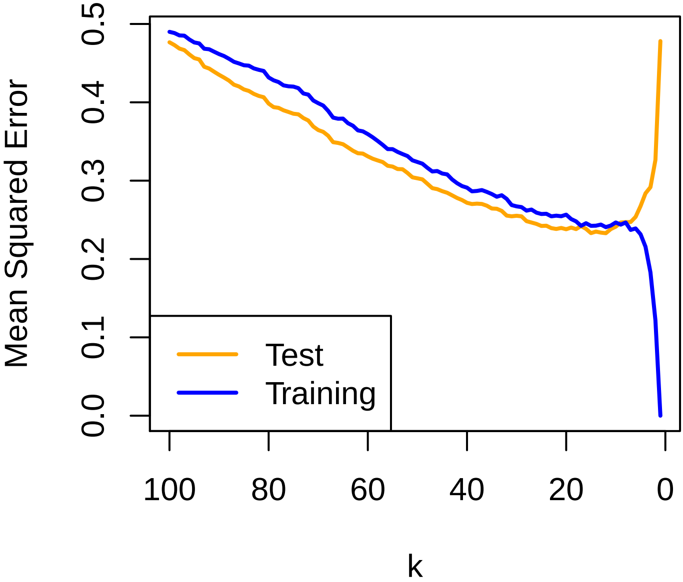


Overfitting

A method is **overfitting** the data when it has a small training MSE but a large test MSE. Using a bigger dataset:



Overfitting via *k*-nearest neighbors



MSE decomposition

Given $Y = f(X) + \epsilon$, where $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$, consider a predictor $\widehat{f}$

Expected MSE for predicting new Y at $X = x$ decomposes as:

$$E[(Y - \widehat{f}(x))^2] = Var(\widehat{f}(x)) + [Bias(\widehat{f}(x))]^2 + \sigma^2$$

Bias-variance tradeoff

Interpretation:

- *Var*$(\widehat{f})$ is the amount of variability in our predictor with respect to the training data.
- *Bias*$(\widehat{f})$ is the systematic error introduced by model approximation.
- σ^2 is *irreducible error*, inherent in the error term ϵ.

Bias-variance tradeoff

Interpretation:

- $Var(\widehat{f})$ is the amount of variability in our predictor with respect to the training data. Increases with increasing model flexibility.
- $Bias(\widehat{f})$ is the systematic error introduced by model approximation. Decreases with increasing model flexibility.
- σ^2 is *irreducible error*, inherent in the error term ϵ. Cannot get rid of this!

If we have a family of flexible regression methods, we should try to balance squared bias and variance.

Summary from today

- Least squares coefficients correspond to minimum of a quadratic surface
- Confidence intervals computed using standard errors of coefficients
- R^2 is a scale-invariant accuracy measure — proportion of variance in Y explained by the model
- Multiple linear regression (many predictors) estimated by solving a linear system — normal equations

Readings in ISL

- Chapter 2 (mostly last lecture)
- Chapter 3 (today and Thursday)

Messing around with R

- R excerpts from slides
- Markdown and R Markdown
- knitr in a knutshell
(`http://kbroman.org/knitr_knutshell/`)