Linear Regression

COS 424/524, SML 302: Fundamentals of Machine Learning

Professor Engelhardt

COS424/524, SML 302

Lecture 7

Prediction

We learned about *classification*, or assigning a class label to an unlabeled sample.

Now we turn to the problem of *prediction*, or assigning a scalar response to a new sample with no observed response.

Prediction is the second of the major goals of machine learning that we are discussing.

Predicting a response variable

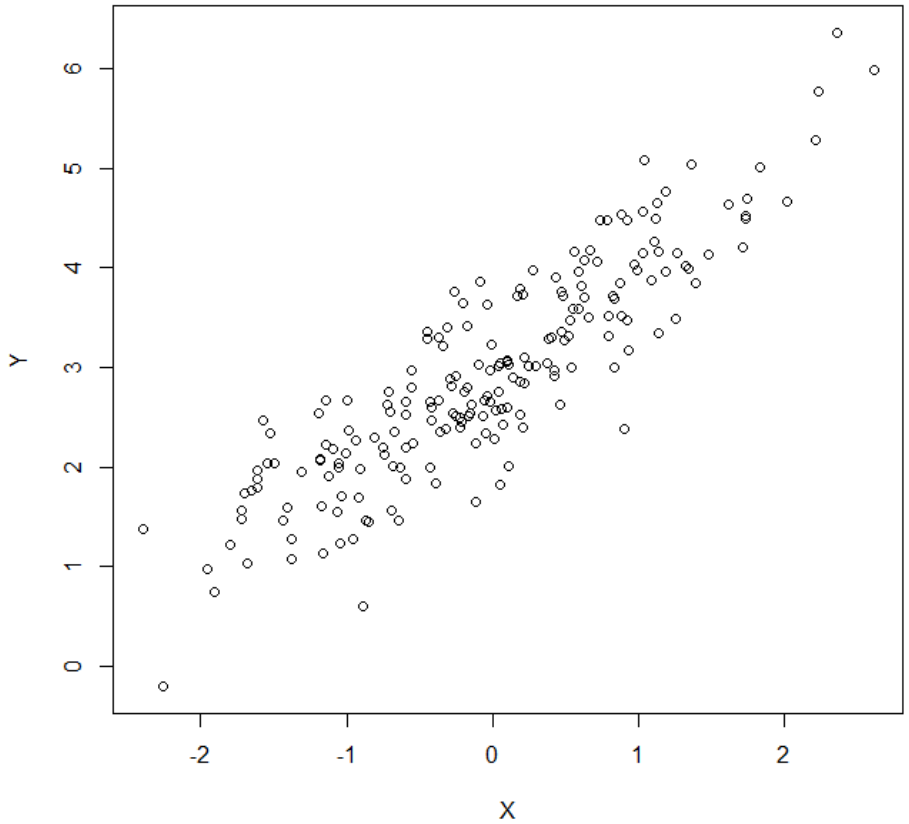
Prediction is a generalization of classification.

A *response variable* can take on many forms

- a value in $\Re$
- a binary variable (binary classification)
- a categorical variable (multiclass classification)
- a value in $\mathcal{Z}^+$
- a value in $(0, 1)$
- many more...

Today we will consider predicting a response in $\Re$ using *linear regression*.

What is linear regression?



Linear regression models capture the linear relationship between a response variable and predictor variables, assuming independent Gaussian noise

What can we use linear regression for?

Given a fitted linear regression model, we may:

- predict the value for y^* for a new observation x^*;
- test for a linear relationship between variables x and y;
- when y is binary or multinomial, we can use this model to classify a new observation x^* (*logistic regression*);
- identify the subset of x variables that are most predictive of y (*feature selection*).

Later in this course, we will learn versions of this model that allow non-Gaussian responses.

Linear regression: examples

Let's think about possible relationships we want to model:

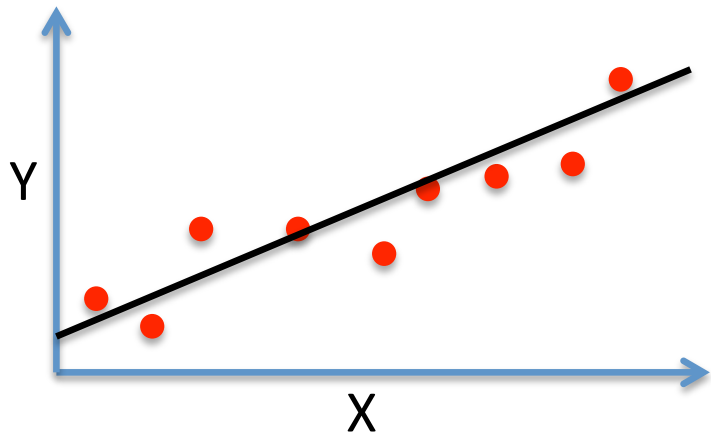
Examples of linear regression models

- $x =$ age and $y =$ height
- $x =$ distance from the shore and $y =$ weight of clams found
- $x =$ gestational age and $y =$ birthweight
- $x =$ cigarette smoker and $y =$ lifespan
- $x =$ presence of a dam and $y =$ fish weight
- $x =$ disposable income and $y =$ total consumption
- $x =$ genotype and $y =$ hip-to-waist ratio

Definitions: univariate linear regression

Sample $i \in 1 : n$, $p = 1$ predictors

- $y_i \in \Re$: response (observed)
- $x_i \in \Re^p$: predictors, covariates, or explanatory variables (observed)
- $\beta \in \Re^p$: coefficients, effects (parameter)
- $\epsilon_i \in \Re$ residual error, noise



$$y_i = \beta x_i + \epsilon_i$$

- Training data consists of n sample pairs (supervised learning):

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}.$$

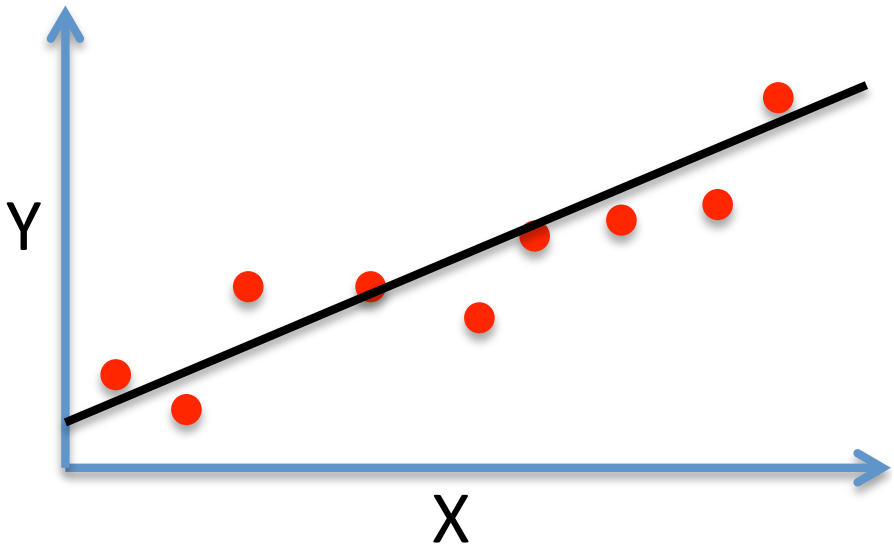
Linear model specification

Gaussian linear regression

A Gaussian linear regression model has the form:

$$y_i = x_i\beta + \epsilon_i$$

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



This is equivalent to:

$$\begin{aligned} (y_i - x_i\beta) &= \epsilon_i \\ (y_i - x_i\beta) \mid \beta, \sigma^2 &\sim \mathcal{N}(0, \sigma^2) \\ y_i \mid x_i, \beta, \sigma^2 &\sim \mathcal{N}(x_i\beta, \sigma^2). \end{aligned}$$

The conditional distribution of $y \mid x$ is Gaussian.

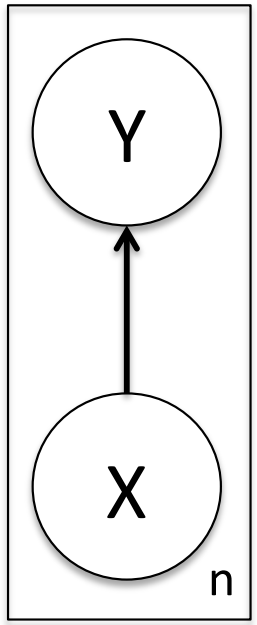
Linear regression as a discriminative model

Gaussian linear regression

A Gaussian linear regression model has the form:

$$y_i \quad = x_i\beta + \epsilon_i$$

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



Writing this out in terms of the graphical model, we have:

$$p(x, y) = p(x)p(y \mid x).$$

We know the conditional distribution of y given x is Gaussian:

$$y_i \mid x_i, \beta, \sigma^2 \sim \mathcal{N}(x_i\beta, \sigma^2).$$

We have not specified the distribution of x; does this matter?

Model fitting and prediction

Fitting a linear regression model

With our model $y_i = x_i\beta + \epsilon_i$ and training data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, we can estimate parameters β and σ^2, giving us $\hat{\beta}$ (coefficient) and $\hat{\sigma}^2$ (residual variance).

Prediction for a new point x^*

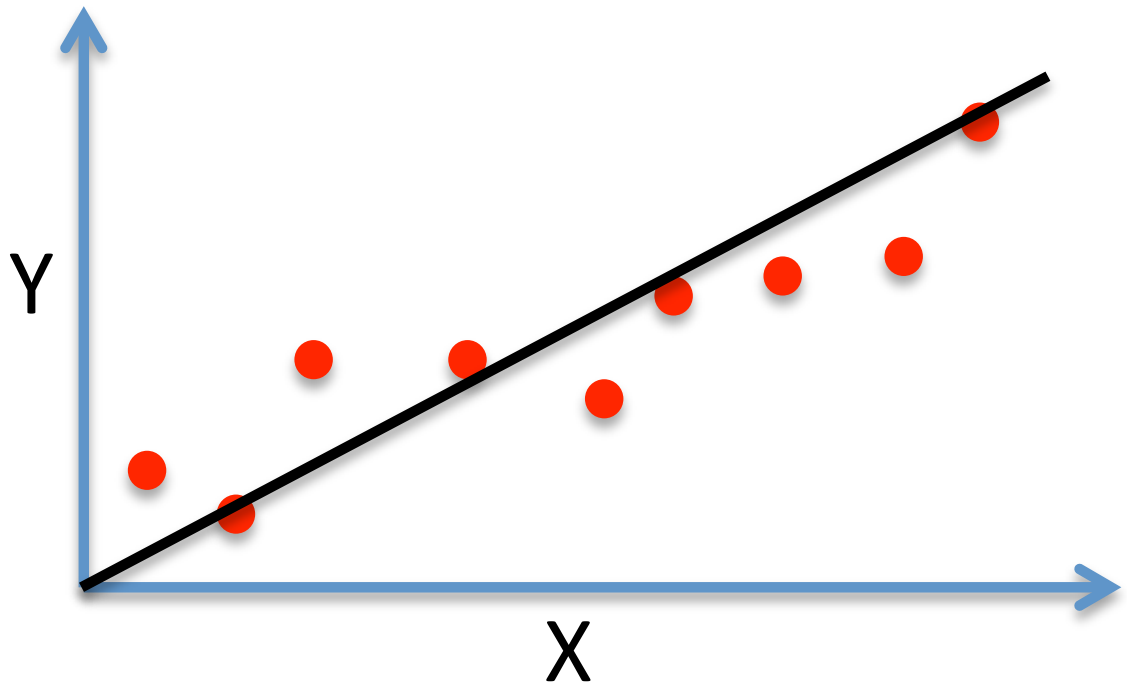
We can predict y^* for a new point x^*, given $\hat{\beta}$:

$$\hat{y}^* = \mathrm{E}[y^*|x^*, \hat{\beta}] = \hat{\beta}x^*.$$

This is not quite right. What else do we need to specify an arbitrary linear relationship between two variables?

Adding the intercept term

With our current model, when $x^* = 0$, $\hat{y}^* = \hat{\beta}x^* = 0$.



What if this is not true in our data (e.g., height at age 0)?

Adding the intercept term

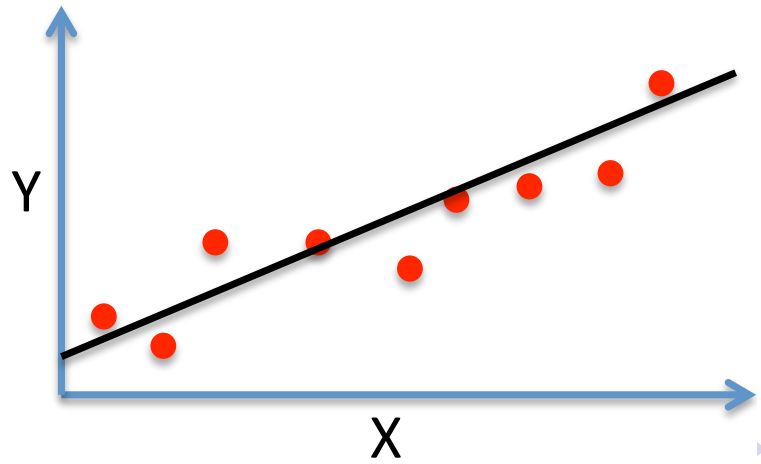
Let's add a y-axis intercept term, β_0.

$$y_i = \beta_0 + \beta x_i + \epsilon_i$$

We can do this in the laziest way possible: add a 1 as the first element of each x_i vector, and a β_0 term as the first element of a β vector:

$$\begin{aligned} \mathbf{x_i} &= [1, x_i]^T \\ \tilde{\beta} &= [\beta_0, \beta]^T \end{aligned}$$

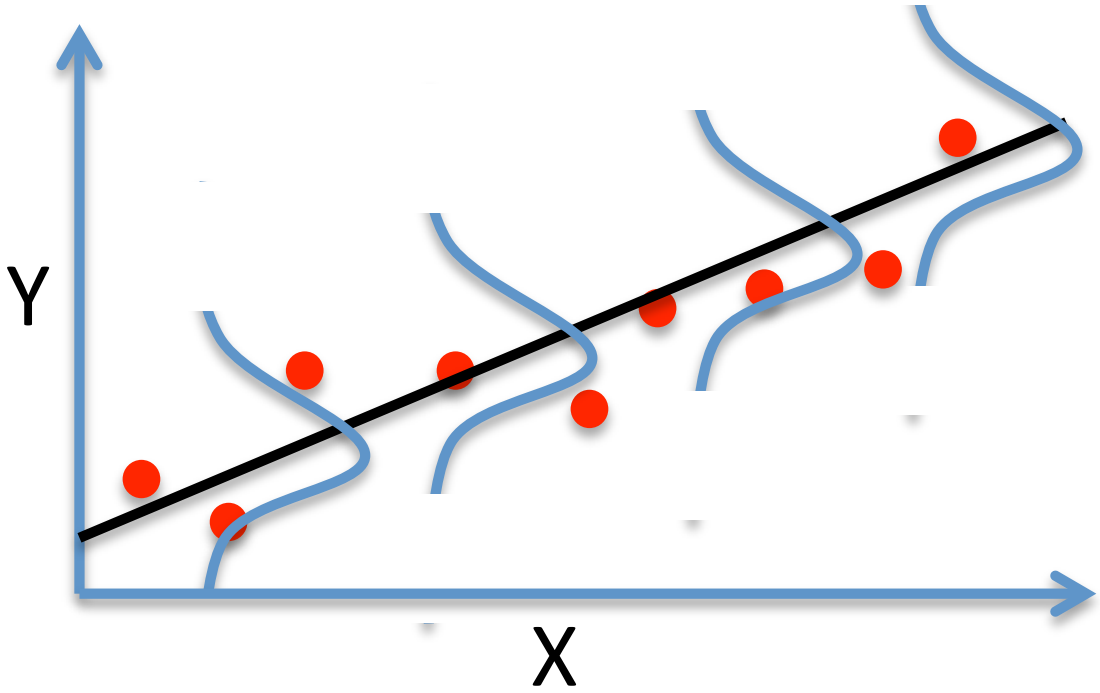
Our model is then: $y_i = \beta_0 + \beta x_i + \epsilon_i = \mathbf{x_i}^T \tilde{\beta} + \epsilon_i$.



Linear regression: conditional distribution

We have defined the conditional probability of the response y:

$$y_i \mid x_i, \beta, \sigma^2 \sim \mathcal{N}(\beta_0 + \mathbf{x_i}\beta, \sigma^2).$$

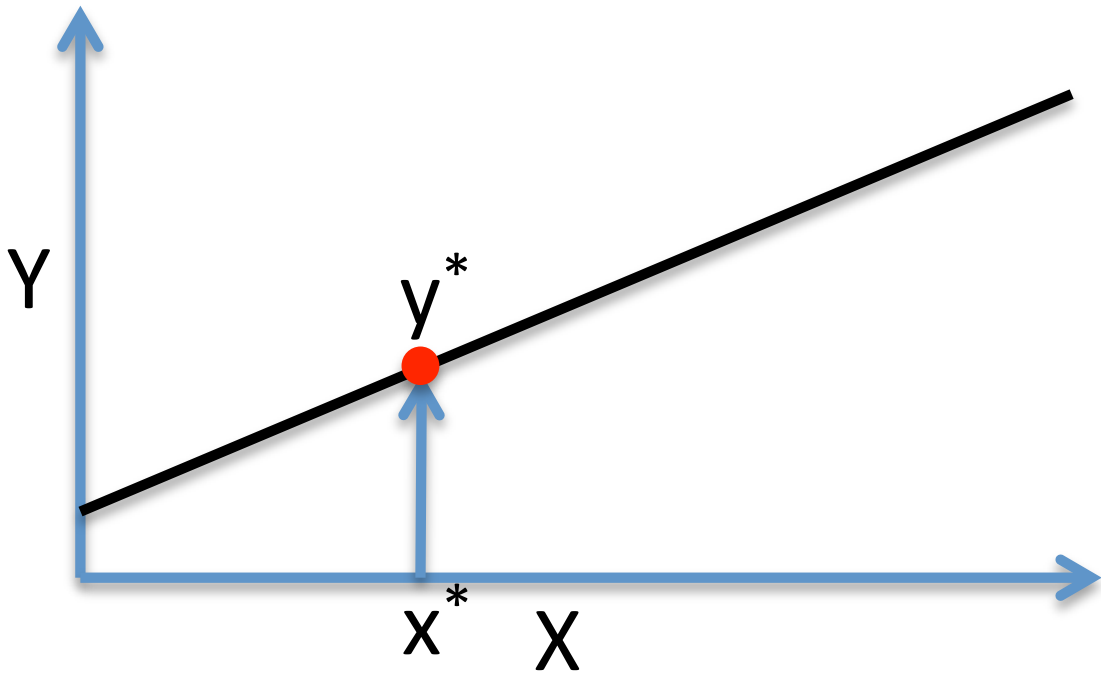


At every value of x_i, the response y_i has a conditionally Gaussian distribution with mean $\beta_0 + \beta x_i$.

Linear regression: prediction

We predict the value of y^* for x^* with the conditional Gaussian mean:

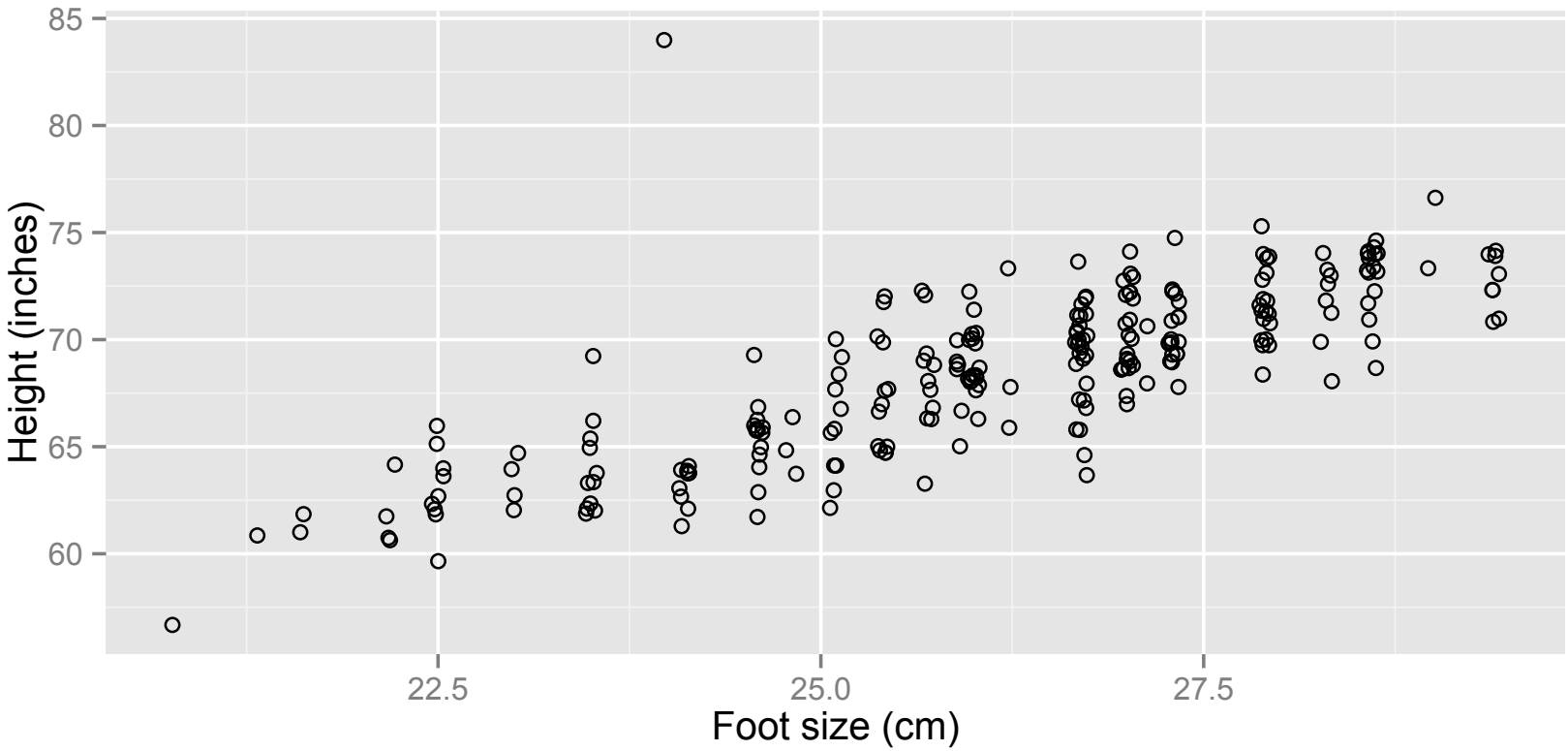
$$y^* = \hat{\beta}_0 + \hat{\beta} x^*.$$



Example: predicting height using shoe size

Let's use our survey data to step through linear regression.

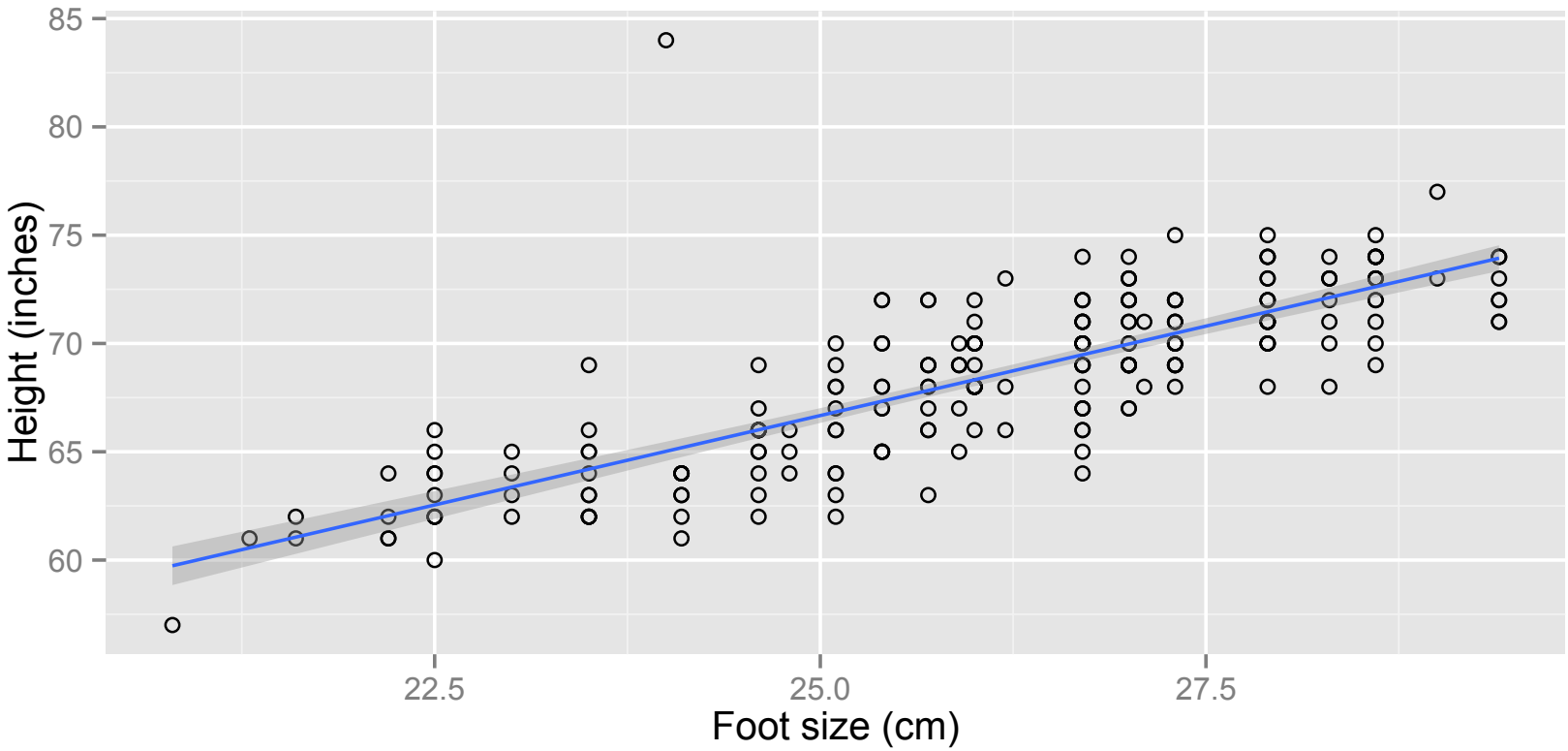
Let y = height (in) and x = shoe size (cm).



Does it look like there is a linear relationship between x and y?

Example: predicting height using shoe size

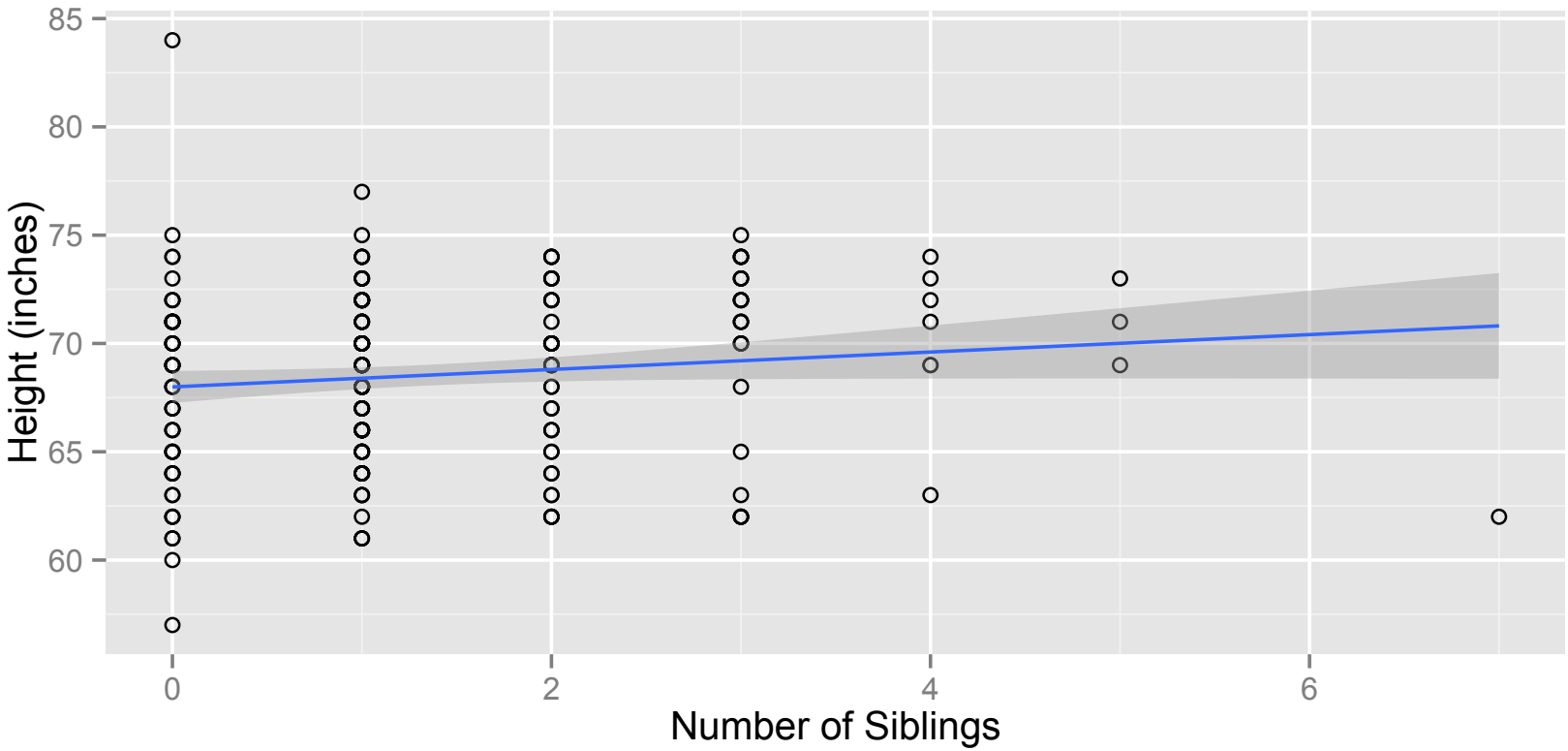
We fit a linear model to these data, and find that $\hat{\beta}_0 = 25.39$, $\hat{\beta} = 1.65$:



Do these parameter estimates support a hypothesis about a linear relationship existing between x and y?

Example: predicting height using number of siblings

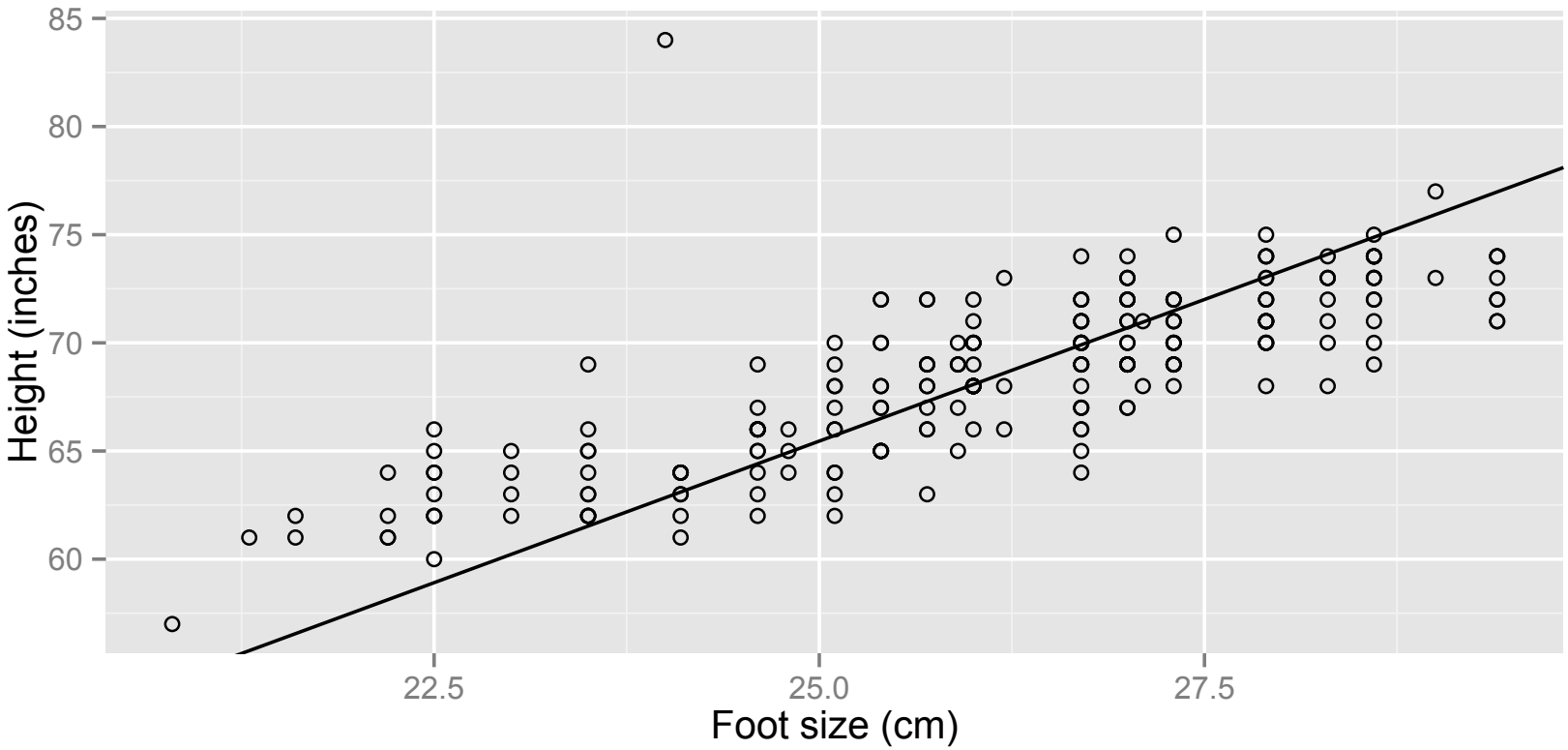
We fit a linear model to these data, and find that $\hat{\beta}_0 = 68.0$, $\hat{\beta} = 0.4$:



Do these parameter estimates support a hypothesis about a linear relationship existing between x and y?

Example: predicting height using shoe size

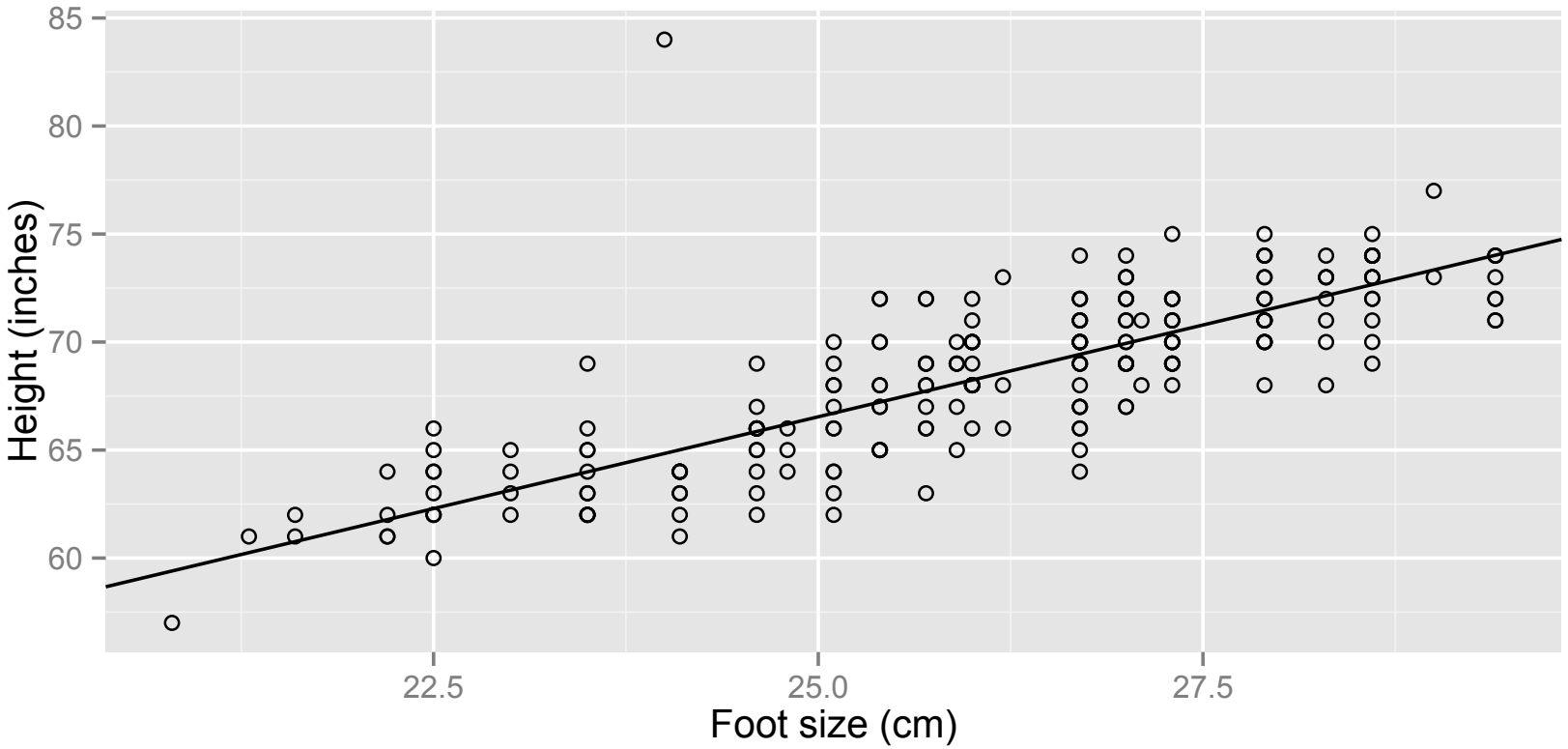
We fit a linear model to $\mathcal{D}$ with no intercept term, and find that $\hat{\beta} = 2.62$ versus $\hat{\beta} = 1.65$



How can we determine how much worse this fit is relative to the model with an intercept term?

Example: predicting height using shoe size

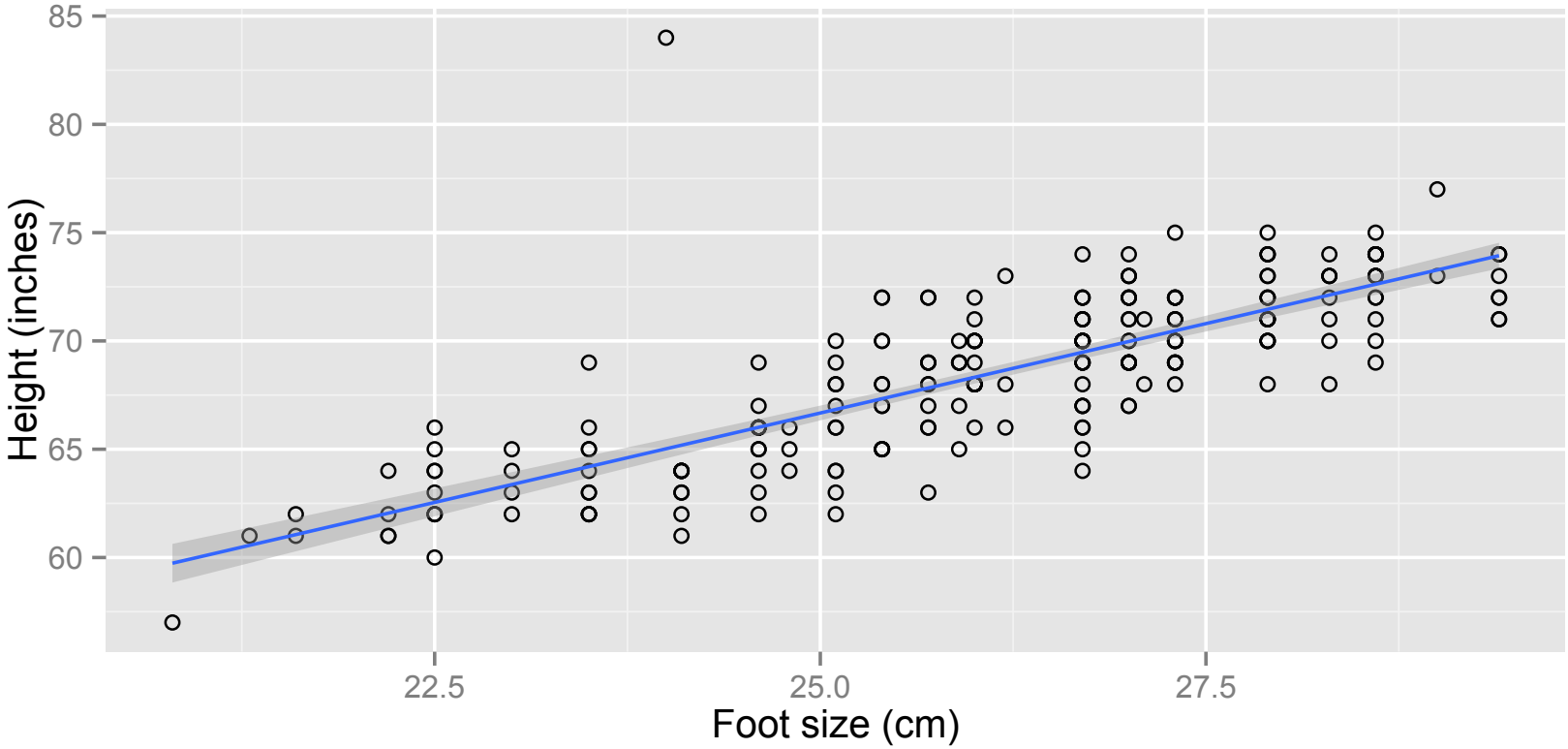
We fit a linear model to $\mathcal{D}$, removing the one outlier point, and find that $\hat{\beta}_0 = 24.1$ and $\hat{\beta} = 1.7$ versus $\hat{\beta}_0 = 25.39$, $\hat{\beta} = 1.65$



How can one point have such a big impact on the fitted regression line?

Example: predicting height using shoe size

Now, say I am a forensics expert, and I find a suspect's shoe print at a crime scene; the shoe is 26 cm long. What do I know about the suspect?



$$\begin{aligned} y^* &= \hat{\beta}_0 + \hat{\beta} x^* \\ 68.3 &= 24.1 + 1.7 \times 26. \end{aligned}$$

Definition: residuals

We define the *residual* as follows (dropping β_0):

$$r_i = y_i - \hat{\beta} x_i = y_i - \hat{y}_i$$
$$\hat{\beta} = \text{our estimated } \beta$$
$$\hat{y}_i = \text{our predicted value}$$

- These r_i are called the *residuals*.
- For each sample, the error is the same as the residual: $\epsilon_i = y_i - \hat{\beta} x_i = r_i$.
- From our definition of the error term

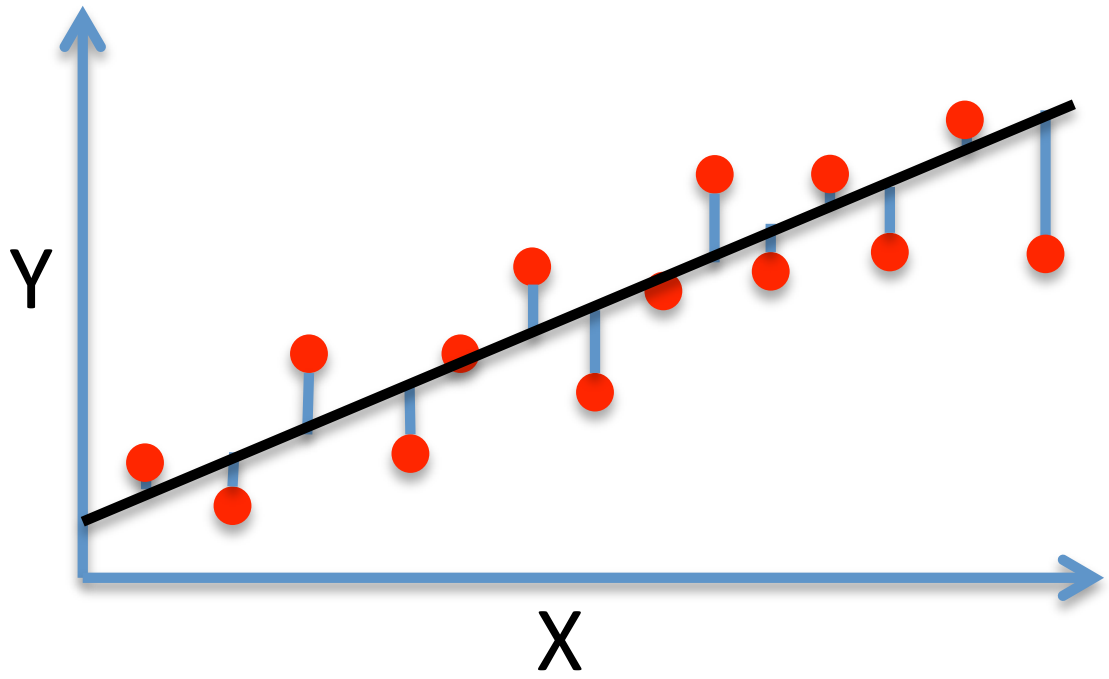
$$\mathbb{E}[r_i] = \mathbb{E}[\epsilon_i] = \mathbb{E}[y_i - \hat{\beta} x_i] = 0$$

- I.e., the expected residual is zero for the Gaussian model.

Linear regression: residuals

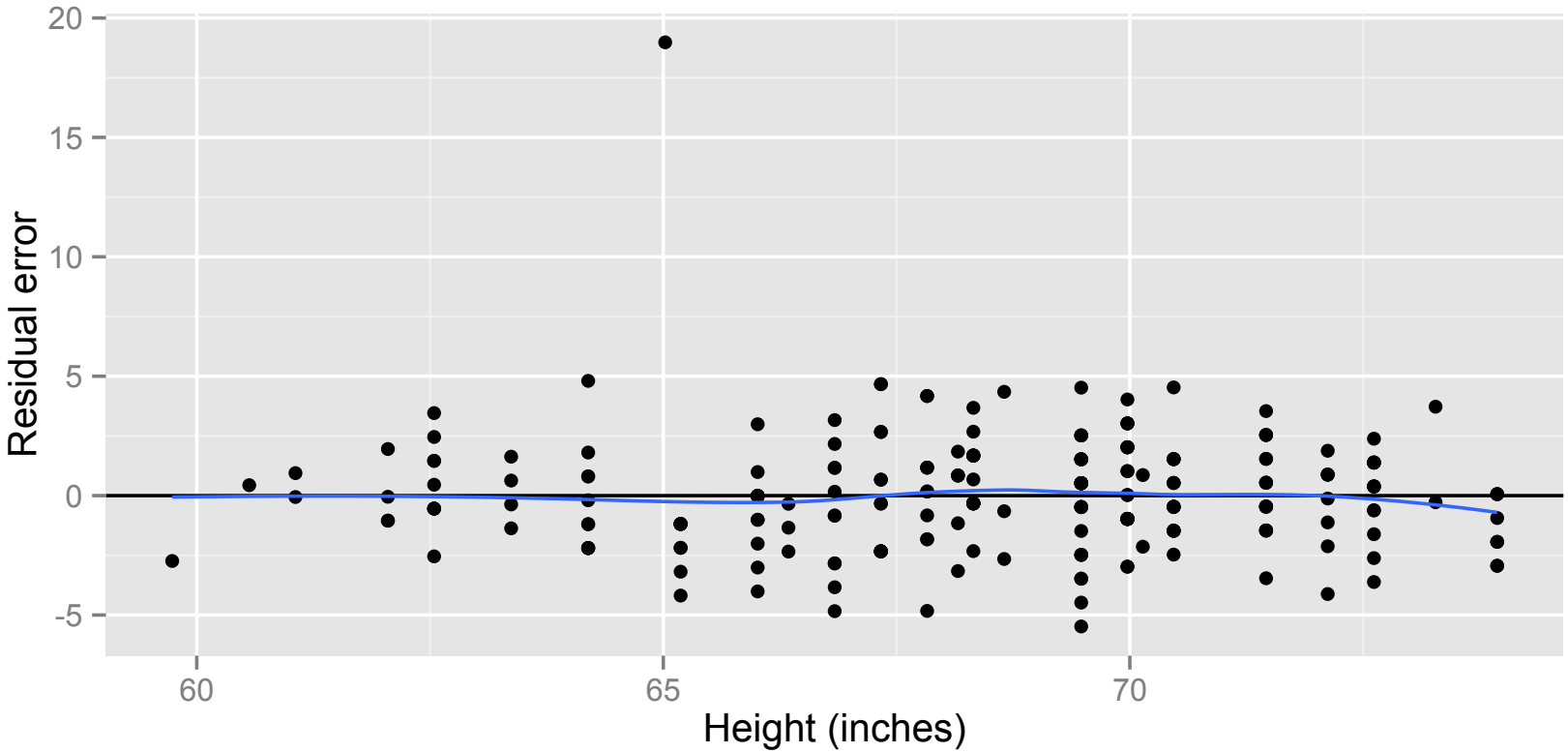
For a fitted linear regression model and a data set $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, I can compute the residuals:

$$r_i = y_i - \hat{\beta} x_i = y_i - \hat{y}_i$$



Example: predicting height using shoe size

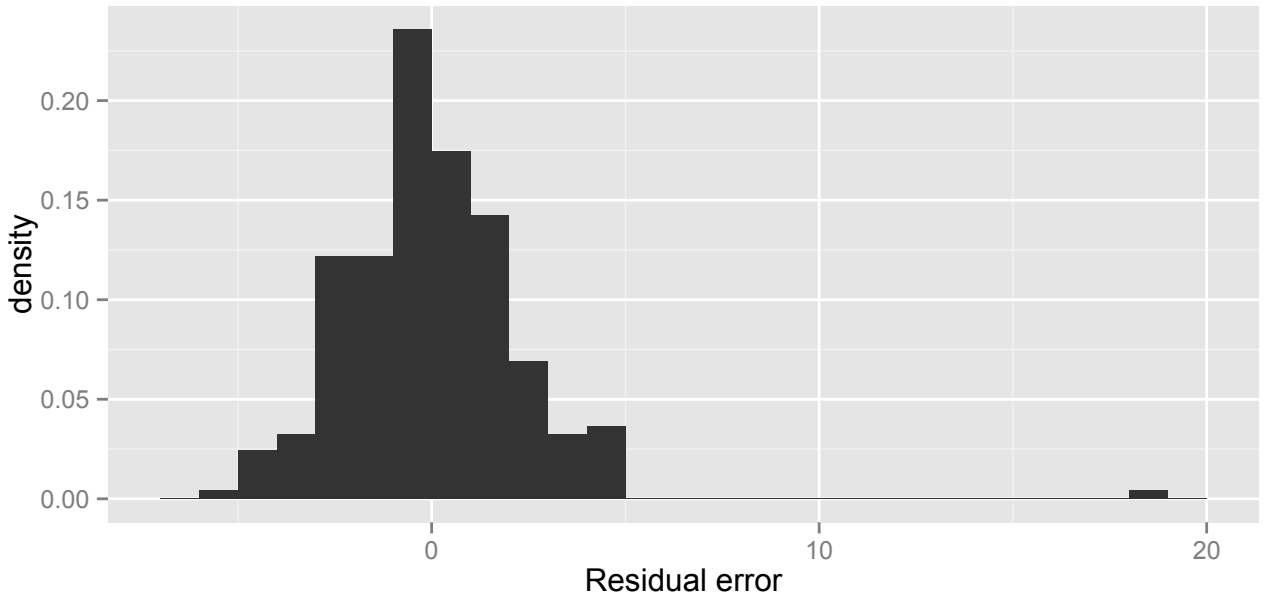
We can plot the residuals of the fitted model:



Are these Gaussian like the model assumes they are?

Example: predicting height using shoe size

We can examine the residuals of the fitted model:



These look approximately Gaussian with zero mean, with one outlier.

Evaluating the fit of the regression model to data I

How do we quantify the quality of the predicted response values?

There are a number of metrics, all of which are a function of the residuals:

- Residual Sum of Squares (RSS):

$$\text{RSS}(\hat{\beta}, \mathcal{D}) = \sum_{i=1}^{n} (y_i - \hat{\beta}^T x_i)^2$$

- RSS: total squared difference between predicted and true response values; lower values indicate better fit.

Evaluating the fit of the regression model to data II

How do we quantify the quality of the predicted response values?

There are a number of metrics, all of which are a function of the residuals:

- Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\text{RSS}(\hat{\beta}, \mathcal{D})$$

- Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\text{MSE}}$$

- MSE, RMSE: (root) average squared difference between predicted and true response values; lower values indicate better fit.

Evaluating the fit of the regression model to data III

How do we quantify the quality of the predicted response values?

There are a number of metrics, all of which are a function of the residuals:

- Coefficient of determination (r^2) :

$$r^2 = 1 - \frac{RSS(\hat{\beta}, \mathcal{D})}{\sum_{i=1}^{n} \left(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i\right)^2}$$

- r^2: correlation between predicted and true response values; higher values indicate better predictions.

Evaluating the fit of our model: example

On our shoe size, height data:

	linear model	no intercept	remove outlier
RSS	815.3	1188.3	453.9
MSE	6.42	9.36	3.60
RMSE	2.53	3.06	1.90
r^2	0.60	0.42	0.75

Multiple predictors: multivariate linear regression

What happens if we have more than one predictor, or feature, that we can use to predict response y?

Examples of multivariate linear regression models

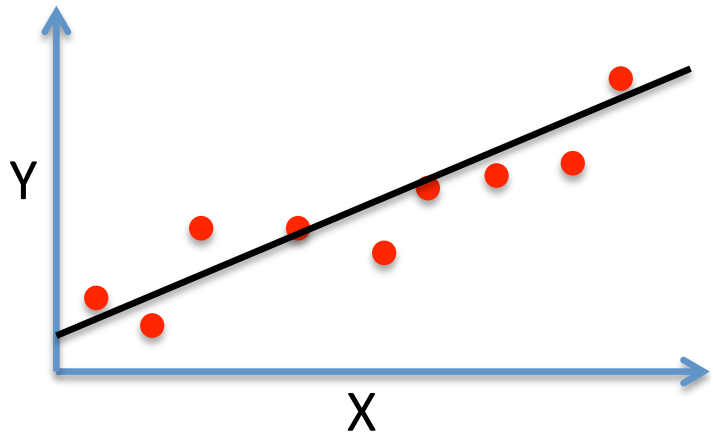
- x = age, gender and y = height
- x = distance from the shore, depth of water and y = weight of clams
- x = gestational age, mother's age and y = birthweight
- x = cigarette smoker, BMI and y = lifespan
- x = presence of a dam, water temperature and y = fish weight
- x = disposable income, education and y = total consumption
- x = genotype at 20 million genomic loci and y = hip-to-waist ratio

We are increasing the number of features, not the number of samples.

Definitions: Multivariate regression

For sample $i \in 1 : n$, p predictors

- $y_i \in \Re$: response (observed)
- $\mathbf{x_i} \in \Re^{\mathbf{p}}$: predictors, covariates, or explanatory variables (observed)
- $\beta \in \Re^{\mathbf{p}}$: coefficients, effects (parameter)
- $\epsilon_i \in \Re$ residual error, noise



Univariate versus multivariate regression

- *Univariate regression* $p = 1$, a single covariate
- *Multivariate regression* $p > 1$, multiple covariates

Linear model specification

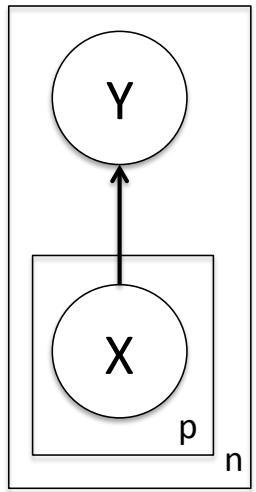
For p predictors and n samples, we define multivariate linear regression.

Gaussian multivariate linear regression

A Gaussian linear regression model has the form, for a single sample (x, y):

$$\begin{aligned} y &= \mathbf{x}^{\mathbf{T}}\beta + \epsilon \\ &= \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p + \epsilon \end{aligned}$$

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



This is equivalent to $y \mid \mathbf{x}, \beta, \sigma^2 \sim \mathcal{N}(\mathbf{x}^{\mathbf{T}}\beta, \sigma^2)$.

Note the v-structure in the model. What makes this tractable for large p?

Multivariate regression assumptions

These models assume over-simplified data:

- predictors x are treated as fixed value RVs. We do not care about their distribution.
- y is a weighted linear combination of the x values
- the variance term is not a function of x (*homoskedasticity*)
- the residual errors are independent
- the predictors are independent

Linear regression, even with these assumptions, is one of our most important data analysis tools.

Parameter estimation in linear regression

Let's discuss how to estimate the coefficients β, β_0 in the univariate model, with data set $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$

First, we will try to derive the maximum likelihood estimate (MLE); recall:

- write the log likelihood
- differentiate with respect to β
- set equal to 0 and solve for β.

Recall: why is the parameter value at the 0 point of the derivative the parameter MLE?

MLE parameter estimation in linear regression

The likelihood is written as a Gaussian conditional distribution:

$$y \mid x, \beta, \sigma^2 \sim \prod_{i=1}^{n} \mathcal{N}(x_i\beta, \sigma^2).$$

Log likelihood for univariate linear regression

$$\begin{aligned}
\ell(\beta; \mathcal{D}) &= \log \prod_{i=1}^{n} \left[\left(\frac{1}{2\pi\sigma^2} \right)^{1/2} \exp \left\{ \frac{-1}{2\sigma^2} (y_i - \beta x_i)^2 \right\} \right] \\
&= \sum_{i=1}^{n} \log \left[\left(\frac{1}{2\pi\sigma^2} \right)^{1/2} \exp \left\{ \frac{-1}{2\sigma^2} (y_i - \beta x_i)^2 \right\} \right] \\
&= -\frac{n}{2} \log (2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta x_i)^2 \\
&= -\frac{1}{2\sigma^2} \mathrm{RSS}(\beta, \mathcal{D}) + c
\end{aligned}$$

Parameter estimation in linear regression

Let's return to the general multivariate case, i.e. $p > 1$, so each data point x_i is a p-dimensional vector. Everything that follows holds for $p = 1$.

We derive the MLE estimate for β as follows:

$$\begin{aligned} \frac{\partial\left[-\ell(\beta;\mathcal{D})\right]}{\partial\beta} &= \frac{\partial[\frac{1}{2}(\mathbf{y}-\mathbf{X}\beta)^{\mathbf{T}}(\mathbf{y}-\mathbf{X}\beta)]}{\partial\beta} \\ &= X^T X\beta - \mathbf{X^T y} \end{aligned}$$

Setting this equation to zero and solving for β, we have

$$\hat{\beta}_{MLE} = (X^T X)^{-1} X^T \mathbf{y}$$

The normal equation

The normal equation

$$\hat{\beta}_{MLE} = (X^T X)^{-1} X^T \mathbf{y}$$

Here, $X^T X$ is a $p \times p$ matrix; $X^T \mathbf{y}$ is a p-vector.

This is a wonderful result: If $X^T X$ is invertible, then we have an efficient, closed form way to compute the MLE for the regression coefficients (at the cost of a matrix inversion).

How does an intercept affect this equation?

Normal equation

If $X^T X$ is non-invertible (singular), then we cannot compute the MLE of β using the normal equation.

When is $X^T X$ singular?

Normal equation

If $X^T X$ is non-invertible (singular), then we cannot compute the MLE of β using the normal equation.

When is $X^T X$ singular?

$X^T X$ is singular when it does not have full column rank, e.g.,:

- the number of samples n is smaller than the number of predictors p
- the predictors are well correlated

What approaches to parameter estimation do not require full column rank?

Least mean squares (LMS) algorithm

Another way to solve for β is to minimize the residual sum of squares using iterative gradient descent methods.

$$Cost\,(\beta) = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

$$\frac{\partial Cost(\beta)}{\partial \beta} = -\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)x_i$$

Least mean squares (LMS) algorithm

LMS algorithm

Then we can use gradient descent to find the estimate for β

$$\beta^{(t+1)} = \beta^{(t)} + \rho \sum_{i=1}^{n} \left(y_i - {\beta^{(t)}}^T x_i \right) x_i$$

where ρ is the step size.

Note here the summation is from 1 to n, which means we have to iterate through all of the n data instances before we update $\beta^{(t+1)}$.

This is called a *batch method*: each parameter update uses the complete data set.

Online methods

It is often tempting to update our estimate one sample at a time:

- n is too large standard machines to hold in memory;
- large n makes each iteration of gradient descent time consuming;
- the data are still coming into the system.

Methods that process the data one sample at a time are called *online methods* or *stochastic methods [Robbins & Monro 1951]*

Stochastic gradient descent

Let's rewrite the LMS algorithm as:

$$\beta^{(t+1)} = \beta^{(t)} + \rho \left(y_i - {\beta^{(t)}}^T x_i \right) x_i,$$

Stochastic gradient descent

Until convergence, we

1. *randomly pick a sample* $(\mathbf{x_i}, \mathbf{y_i})$ from the data
2. update our estimate of β using only this sample:

$$\beta^{(t+1)} = \beta^{(t)} + \rho \left(y_i - {\beta^{(t)}}^T \mathbf{x_i} \right) \mathbf{x_i},$$

Note that we use all of the p covariates, but only one sample per iteration.

Projecting predictors to a high, non-linear feature space

We can transform the predictors $\mathbf{x}$ using any nonlinear basis function (kernel) in the linear regression model:

$$y = \phi(\mathbf{x})^{\mathbf{T}}\beta + \epsilon$$

This is still a *linear model*: regardless of the form of the basis function $\phi(\cdot)$, parameter β enters the model in a linear way.

Furthermore, the log likelihood is convex with respect to β, and this is true for all $\phi(x)$, so all of the methods proceed identically.

Regression with non-linear relationships

Polynomial functions

We might consider the model:

$$y_i = \phi(\mathbf{x_i})^{\mathbf{T}}\beta + \epsilon$$
$$\phi(x) = [1, x, x^2, \ldots, x^p]^T$$

This allows us to find non-linear (polynomial) relationships between x, y.

Higher-order interactions between predictors

We can include non-additive interactions, e.g., for $x_i = [x_1, x_2]_i$:

$$\phi(x) = [x_1, x_2, x_1x_2, x_1^2x_2, \ldots]$$

Note that we consider interactions among predictors, not among samples.

Summary

A few key points from this discussion:

- Regression models, like other models, must be built thoughtfully:
  - What are the set of predictors to include?
  - Should I include an intercept term?
  - Should I include any non-linear terms?
- With large numbers of samples, online methods might be faster than batch methods.
- With large numbers of predictors, these methods may be unstable
- Optimization methods are an important tool in the ML toolbox

We will discuss regression over the next three lectures.

Additional resources

- Regression is a standard tool in statistics; resources abound
- MLAPA: Chapter 7
- *Elements of Statistical Learning*, Chapter 4
- Metacademy: *Linear Regression*