PSTAT 131/231: Introduction to Statistical Machine Learning

Guo Yu

Lecture 4
Linear Regression

ISL Chapter 3

ESL (for 231 students) Chapter 3.1-3.2, 3.5

Homework 1 is due next Monday, October 11, 2021, 11:59 PM

Quiz

On GauchoSpace

Starting this Friday

You decide when to start the quiz (12:00 PM to 9:00 PM)

Once you start the quiz, you have 20 minutes to finish it

Last time...

Machine Learning vs **Statistical Machine Learning**

Supervised Learning vs **Unsupervised Learning**

Regression vs **Classification**

Training MSE / Error rate vs Test MSE / Error rate

Bias-variance tradeoff

k-NN methods: regression and classification

This week…

Machine Learning vs **Statistical Machine Learning**

Supervised Learning vs **Unsupervised Learning**

Regression vs **Classification**

Linear regression: simple regression and multiple regression

Practical consideration in linear regression

Next week…

Machine Learning vs Statistical Machine Learning

Supervised Learning vs Unsupervised Learning

Regression vs $\boxed{\text{Classification}}$

Logistic regression

Discriminant Analysis

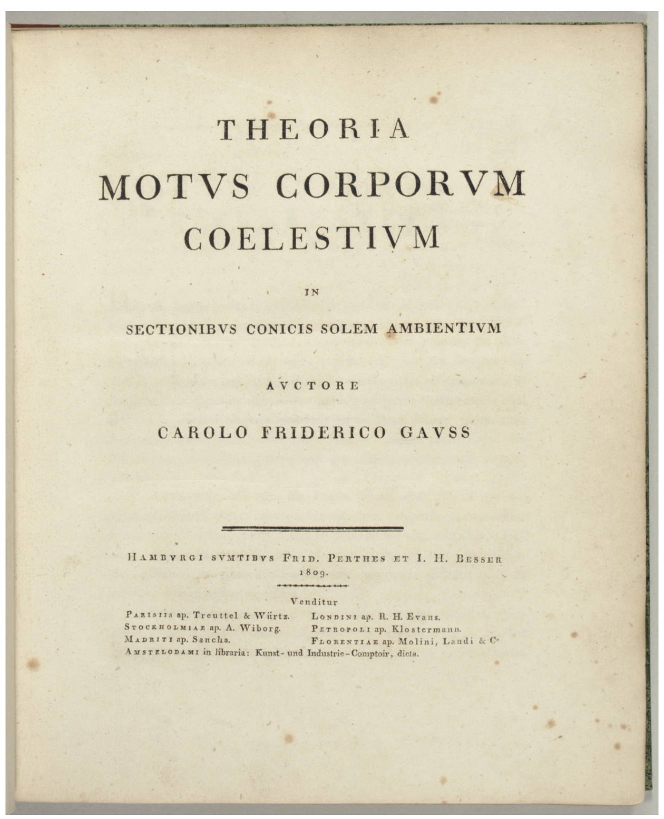
Linear regression

PSTAT 126

Carl Friedrich Gauss in 1795

Simple, interpretable, and often very useful

Many nonlinear methods are direct generalization of linear regression



THEORIA
MOTVS CORPORVM
COELESTIVM
IN
SECTIONIBVS CONICIS SOLEM AMBIENTIVM
AVCTORE
CAROLO FRIDERICO GAVSS

HAMBVRGI SVMTIBVS FRID. PERTHES ET I. H. BESSER
1809.

Venditur
PARISIIS ap. Treuttel & Würtz. LONDINI ap. R. H. Evans.
STOCKHOLMIAE ap. A. Wiborg. PETROPOLI ap. Klostermann.
MADRITI ap. Sancha. FLORENTIAE ap. Molini, Landi & C°
AMSTELODAMI in libraria: Kunst- und Industrie-Comptoir, dicta.

Yes, when I was 18…

(1809): Method of least squares, MLE, Gaussian distribution…

Recap: Bias-variance decomposition

If we take

$$\hat{f}(\mathbf{x}_0) = \mathrm{E}\left[Y \,|\, X = \mathbf{x}_0\right]$$

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

non-random **zero-mean noise**

$$\mathrm{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] = \mathrm{Var}(\hat{f}(\mathbf{x}_0)) + \left[\mathrm{Bias}(\hat{f}(\mathbf{x}_0))\right]^2 + \mathrm{Var}(\varepsilon),$$

Expected test MSE = **Variance** + **Bias**^{2} + **Irreducible error**

$$= \underbrace{\mathrm{E}\left[\left(\hat{f}(\mathbf{x}_0) - \mathrm{E}\hat{f}(\mathbf{x}_0)\right)^2\right] + \left[\mathrm{E}\left[\hat{f}(\mathbf{x}_0)\right] - f(\mathbf{x}_0)\right]^2}_{\textbf{minimized!}} + \mathrm{Var}(\varepsilon)$$

$$= 0 + 0 + \mathrm{Var}(\varepsilon)$$

Recap: Bias-variance decomposition

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

non-random zero-mean noise

The "best" we can do

$$\hat{f}(\mathbf{x}_0) = \mathrm{E}\left[Y \,|\, X = \mathbf{x}_0\right]$$

Unknown in practice!!

Because the joint distribution of (X, Y) is unknown in practice

What should we do? **Make assumptions:**

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon, \quad \mathrm{E}[\varepsilon] = 0, \quad \varepsilon \text{ independent of } (X_1, \ldots, X_p)$$

Then

$$\mathrm{E}[Y \,|\, X = x_0] = \beta_0 + x_{0_1}\beta_1 + \cdots + x_{0_p}\beta_p$$

The conditional expectation of Y is linear

Linear regression

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon$$

The conditional expectation of Y is linear in the parameters

Simple Linear Regression

$$p = 1$$

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

Multiple Regression

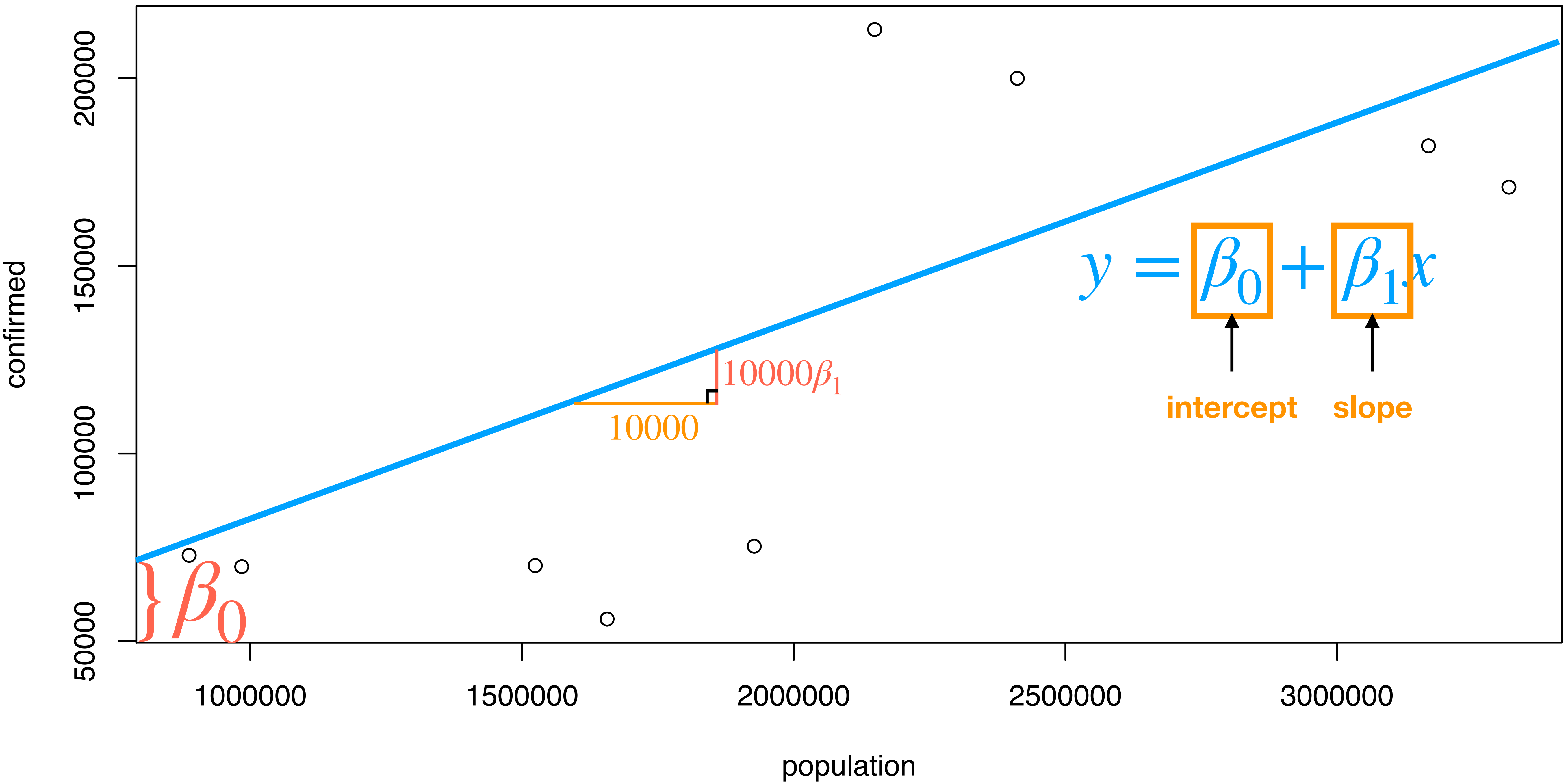
$$p > 1$$

$$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$

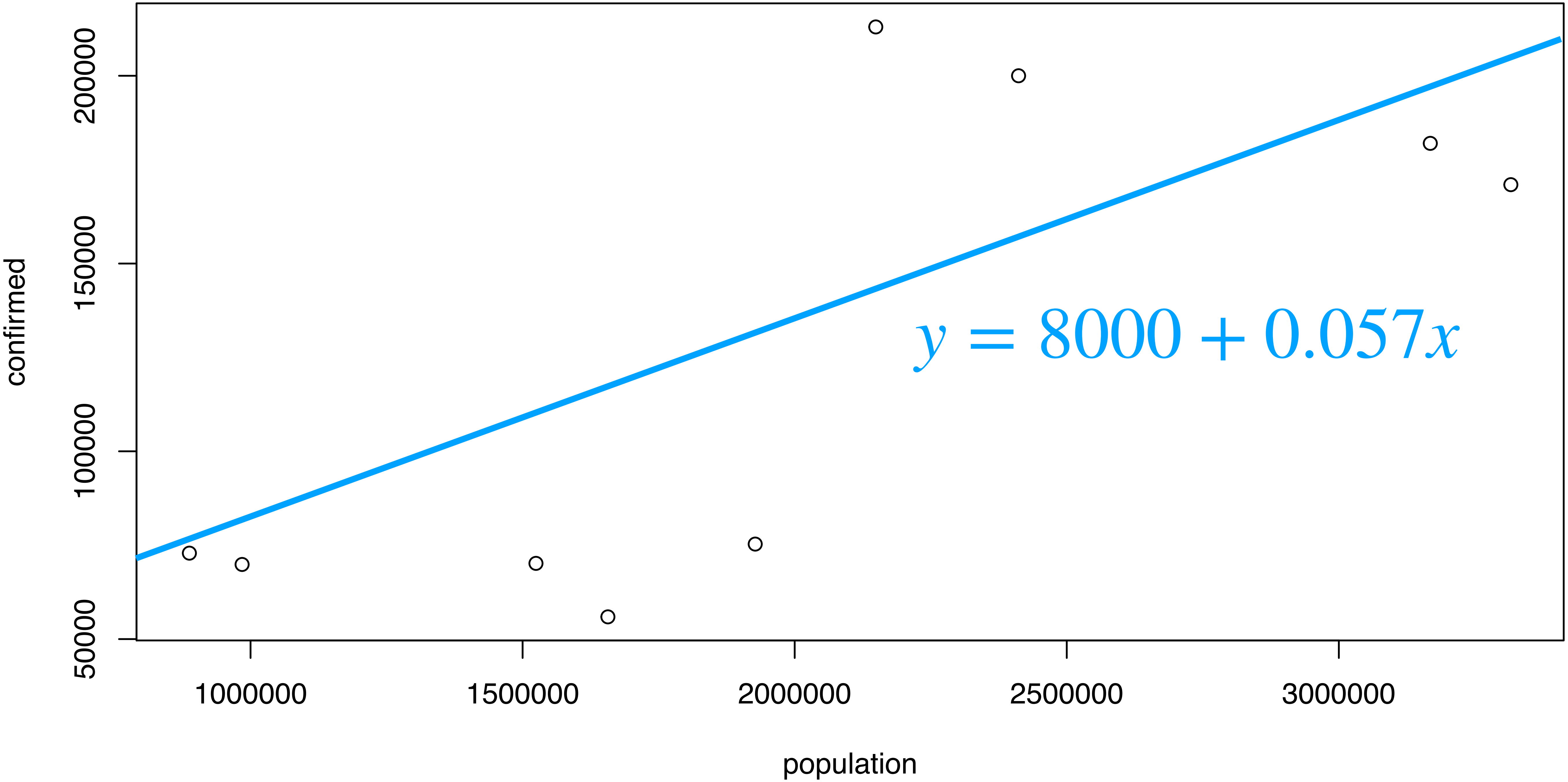
Simple linear regression

$y =$ **confirmed cases** $\quad x =$ **population**



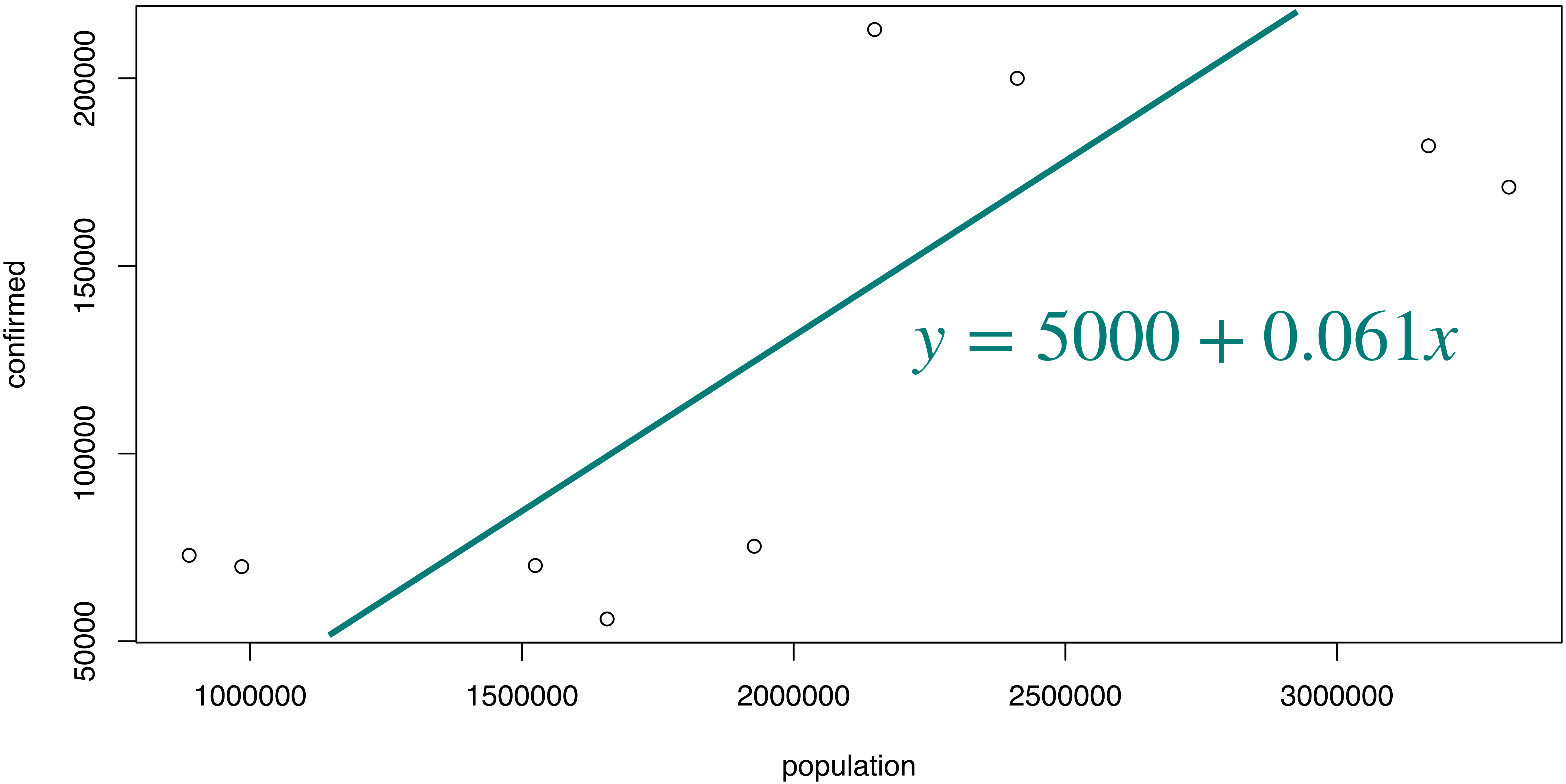
Estimating β_0 and β_1

Different values of β_0 and β_1 give us different lines



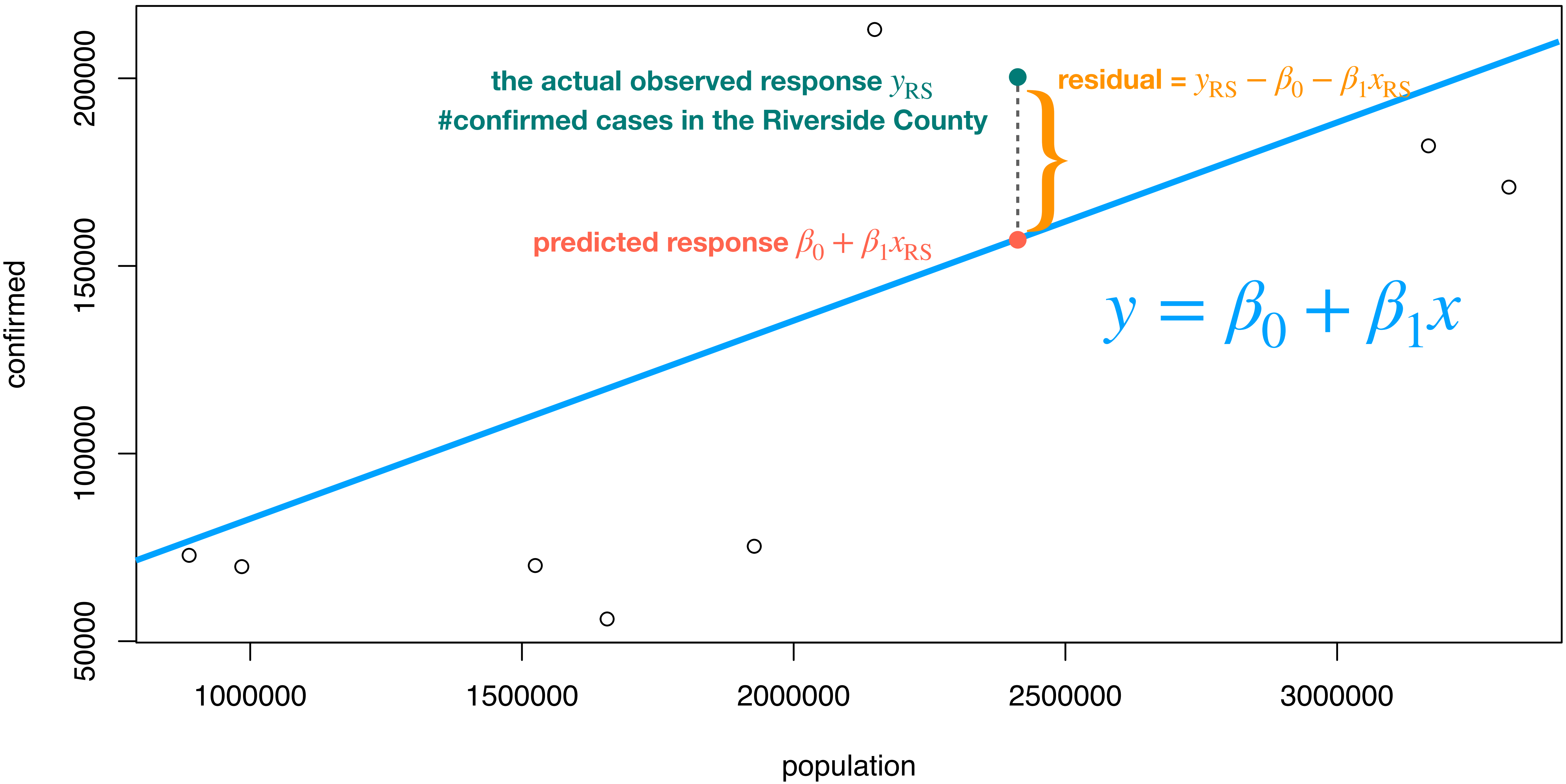
Estimating β_0 and β_1

Different values of β_0 and β_1 give us different lines

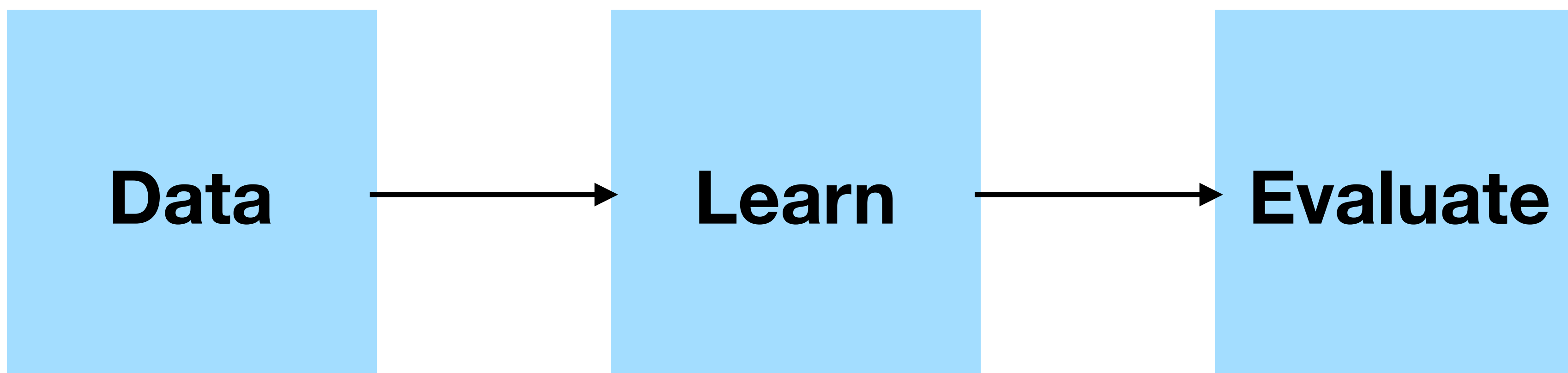


Estimating β_0 and β_1

Which line should we choose?



Simple linear regression



Training set
$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$

$$\hat{\beta}_0, \hat{\beta}_1$$

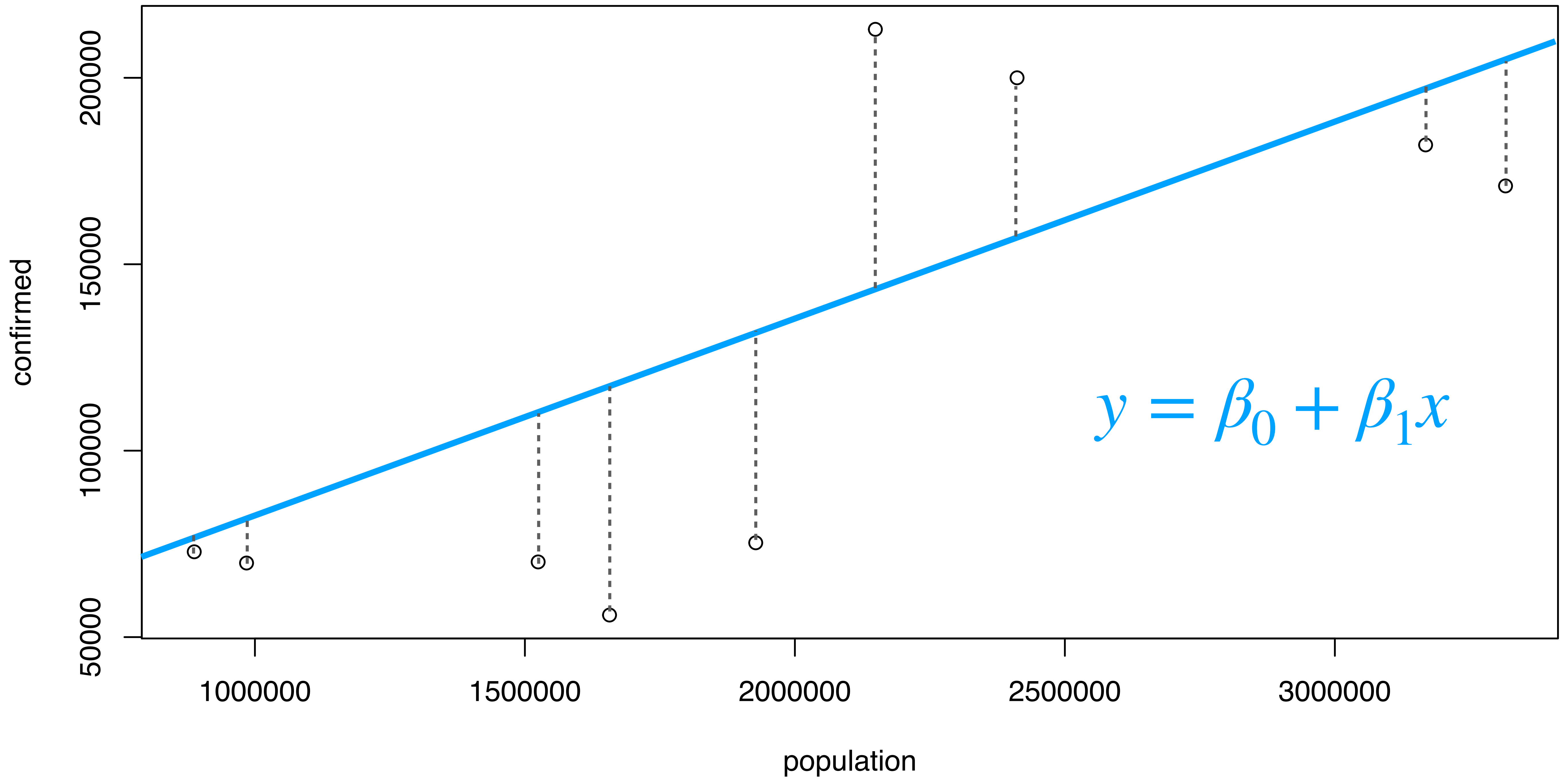
1. Accuracy of the coefficient estimates $\hat{\beta}_0, \hat{\beta}_1$

2. Accuracy of the linear model

Estimating β_0 and β_1

We find the line that best fits the data, by minimizing the sum of squared residuals

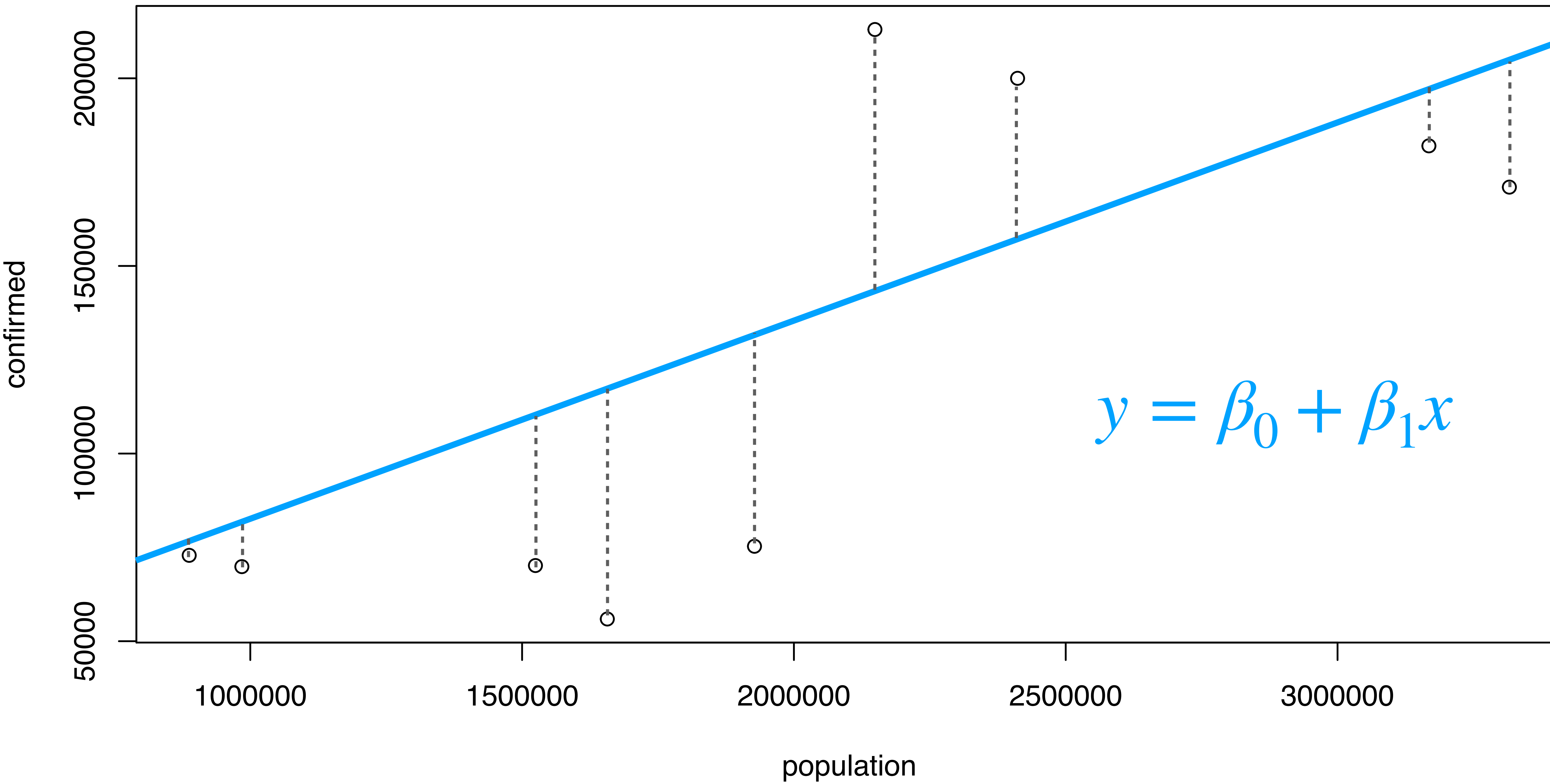
$$\text{SSR} = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$



Estimating β_0 and β_1

We find the line that best fit the data, by minimizing the sum of squared residuals

$$\text{SSR} = \sum_i (\underbrace{y_i - \beta_0 - \beta_1 x_i}_{\text{residual of the } i\text{-th data point}})^2$$

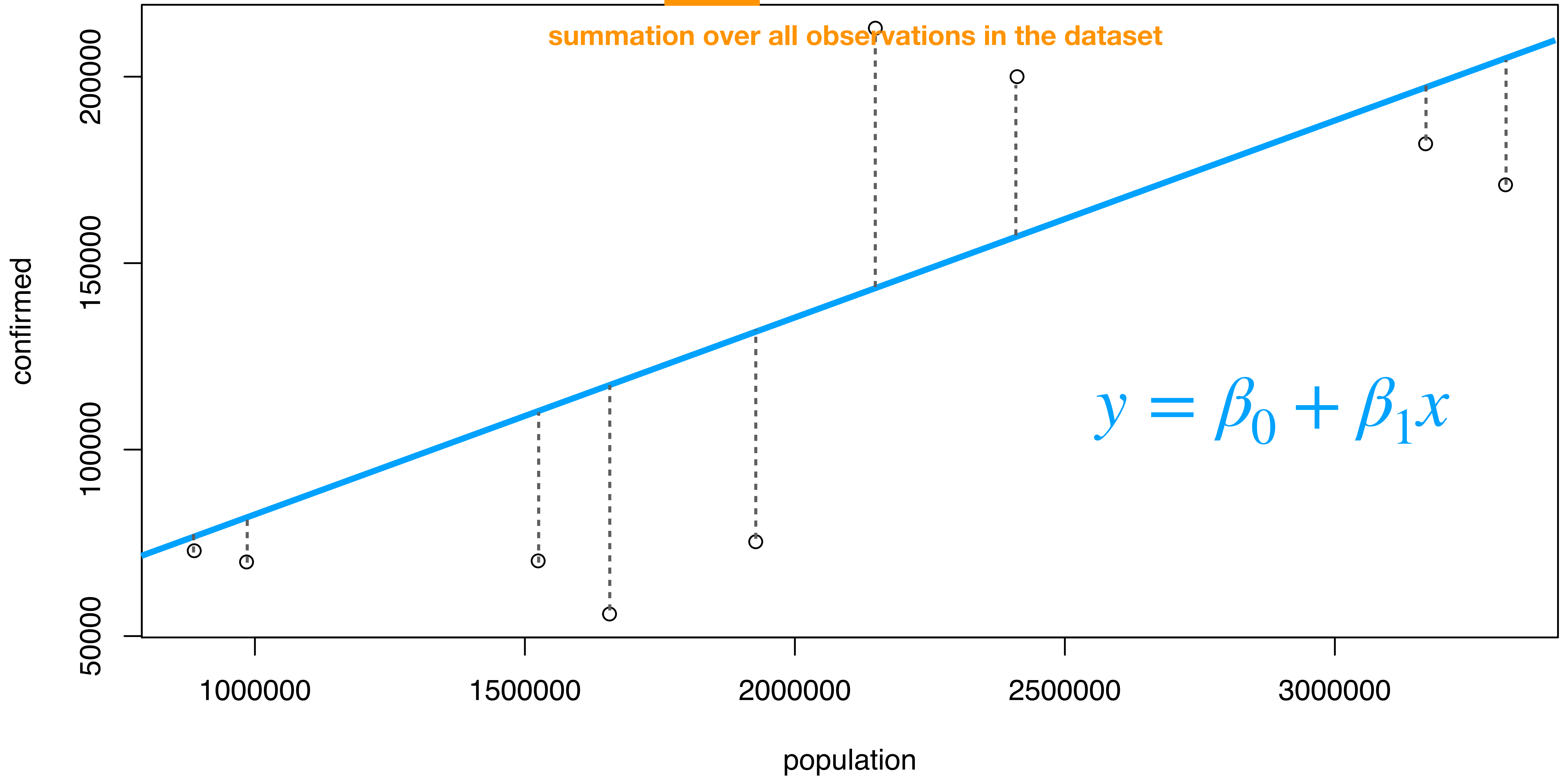


Estimating β_0 and β_1

We find the line that best fit the data, by minimizing the sum of squared residuals

$$\mathrm{SSR} = \boxed{\sum_i}(y_i - \beta_0 - \beta_1 x_i)^2$$

summation over all observations in the dataset



Coefficient estimates

Least-squares coefficient estimates

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

optimal in the sense that they **minimize** the **sum of squared residuals (SSR)**

Accuracy of the coefficient estimates

Least-squares coefficient estimates

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

optimal in the sense that they **minimize** the **sum of squared residuals (SSR)**

Model: $$Y = \beta_0 + X_1\beta_1 + \varepsilon$$

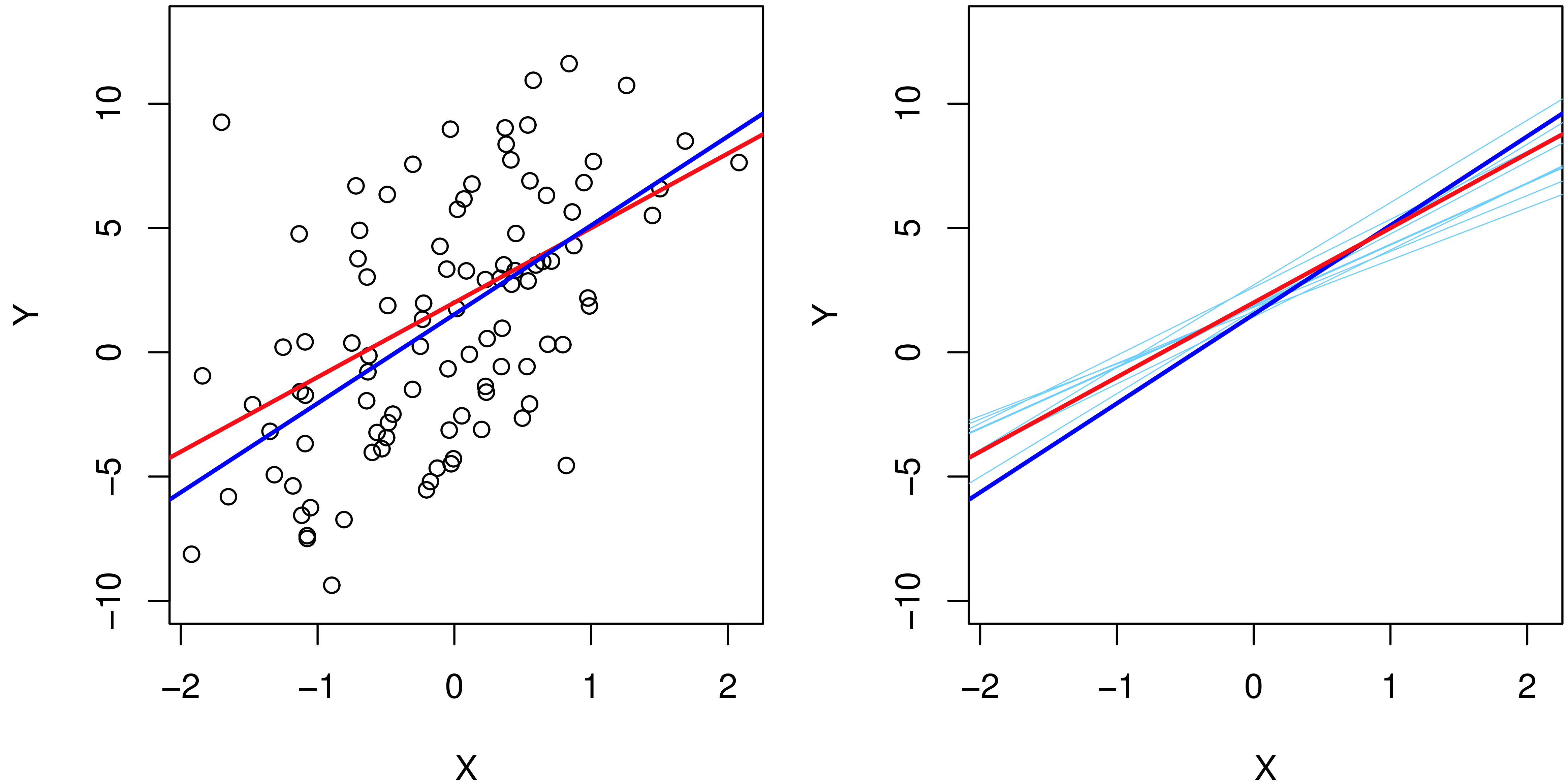
ε: catch-all for what we miss with this simple model

Assumed to have mean zero and independent of everything else

We want: $\hat{\beta}_0 = \beta_0$ and $\hat{\beta}_1 = \beta_1$

But this is **impossible**, since the data are random

Accuracy of the coefficient estimates



$Y = \beta_0 + \beta_1 X$ $\quad$ $Y = \hat{\beta}_0 + \hat{\beta}_1 X$ $\quad$ $Y = \hat{\beta}_0 + \hat{\beta}_1 X$

each computed on the basis of a separate random set of observations

Accuracy of the coefficient estimates

Least-squares coefficient estimates

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables themselves

need to assess how “close” $\hat{\beta}_0$ and $\hat{\beta}_1$ are to β_0 and β_1, respectively

In terms of **bias** and **variance**

Bias of the coefficient estimates

Least-squares coefficient estimates

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

It can be shown that

$$\mathrm{E}(\hat{\beta}_1) = \beta_1 \qquad \mathrm{E}(\hat{\beta}_0) = \beta_0$$

Recall from the bias definition

$$\mathrm{Bias}(\hat{\beta}_1) = \mathrm{E}(\hat{\beta}_1) - \beta_1 = 0 \qquad \mathrm{Bias}(\hat{\beta}_0) = \mathrm{E}(\hat{\beta}_0) - \beta_0 = 0$$

$\hat{\beta}_1$ and $\hat{\beta}_0$ are **unbiased estimator** of β_1 and β_0, respectively

Variance of the coefficient estimates

Least-squares coefficient estimates

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

It can be shown that

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

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

where $\sigma^2 = \mathrm{Var}(\varepsilon)$

Variance of the coefficient estimates

Least-squares coefficient estimates

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

In practice, σ^2 is usually unknown **estimate:** $\hat{\sigma}^2 = \frac{\mathrm{SSR}}{n-2}$

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

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

CI for the coefficients

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

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

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

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

$$\left[\hat{\beta}_1 - 2\,\widehat{\mathrm{SE}}(\hat{\beta}_1), \hat{\beta}_1 + 2\,\widehat{\mathrm{SE}}(\hat{\beta}_1)\right]$$

95% confidence interval (CI) for β_1

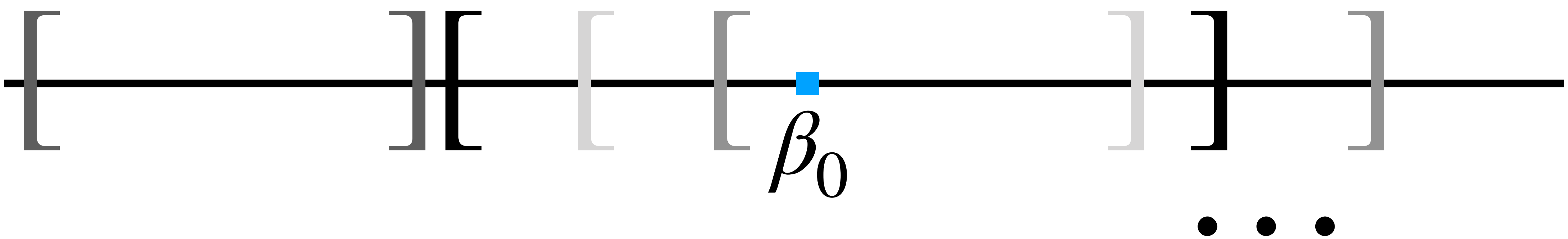
$$\left[\hat{\beta}_0 - 2\,\widehat{\mathrm{SE}}(\hat{\beta}_0), \hat{\beta}_0 + 2\,\widehat{\mathrm{SE}}(\hat{\beta}_0)\right]$$

95% confidence interval (CI) for β_0

There is approximately a 95% chance that the interval will **contain** the true value of β_0 (or β_1)

CI for the coefficients

There is approximately a 95% chance that the interval will **contain** the true value of β_0 (or β_1)



Every data set → a CI

Hypothesis tests on the coefficients

	In words	Mathematically
H_0	**There is no relationship between X and Y**	$\beta_1 = 0$
H_a	**There is some relationship between X and Y**	$\beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1}{\widehat{\mathrm{SE}}(\hat{\beta}_1)}$$

A large value of $|t|$ tends to reject the null hypothesis

t **follows a t-distribution of degrees of freedom $n-2$ under the null**

Accuracy of the model

Residual standard error (RSE)

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{SSR}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

an absolute measure of lack of fit

R squared (R^2)

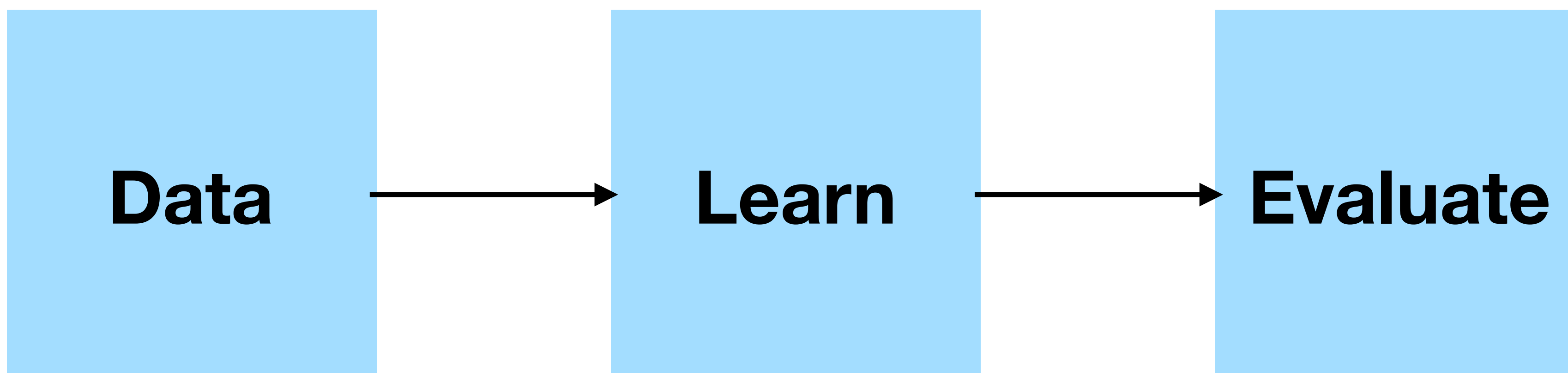
$$\mathrm{R}^2 = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \in [0,1]$$

proportion of the variability in Y that can be explained using X

Both **RSE** and R^2 favor **flexible** methods, which may overfit the data!!

Multiple linear regression

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon$$



Training set

$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$

$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$

Many questions to ask!

$\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$

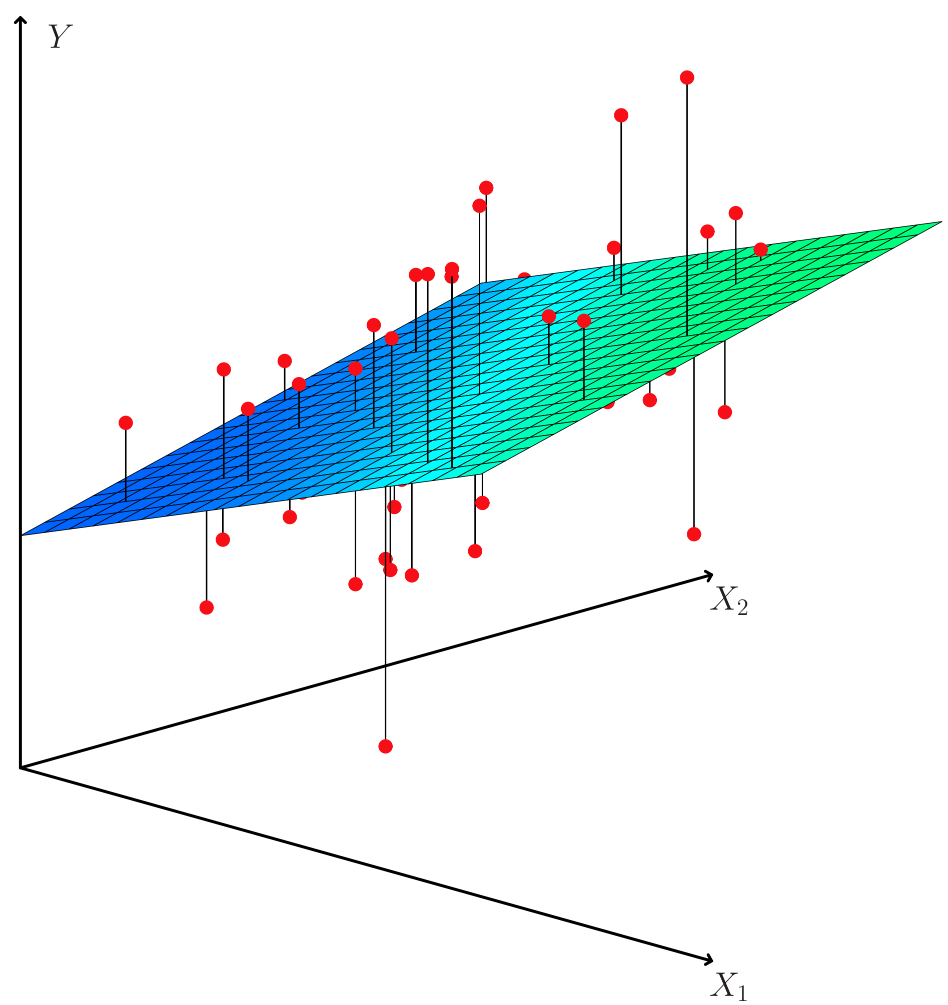
Coefficient estimates

Least-squares coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$

optimal in the sense that they **minimize** the **sum of squared residuals (SSR)**

$$\mathrm{RSS} = \sum_{i=1}^{n} \underline{(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_p x_{ip})}^2$$

ith residual: $y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij}$



Accuracy of the coefficient estimates

Least-squares coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$

random variables themselves

need to assess how “close” $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are to $\beta_0, \beta_1, \ldots, \beta_p$, respectively

In terms of **bias** and **variance**

A few important questions

Q1: Is at least one of the predictors $X_1, \ldots, X_p$ useful in predicting Y?

Q2: Do all the predictors help to explain Y, or is only a subset of the predictors useful?

Q3: How well does the model fit the data?

Q1: Relationship between response and predictors

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

$$H_a : \textbf{at least one } \beta_j \textbf{ is non-zero}$$

We use the **F-statistics**

$$F = \frac{(\mathrm{SST} - \mathrm{SSR})/p}{\mathrm{SSR}/(n-p-1)}$$

$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$\mathrm{SSR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

When H_0 holds,

$$(\mathrm{SST} - \mathrm{SSR})/p \approx \sigma^2$$
$$\mathrm{SSR}/(n-p-1) \approx \sigma^2$$
$$\longrightarrow \quad F \approx 1$$

A large value of F favors H_a

Q2: Deciding on Important Variables

Best subset selection: compute least squares fit for all possible subsets of predictors, and then choose the "best"

"best" in terms of some criteria (Mallow's C_p, AIC, BIC, adjusted R^2, etc.)

impossible for large p: 2^p models in total (~ one billion model when $p = 40$)

Forward selection: start with no variables in the model, add variables one-by-one. Greedy approach.

Backward selection: start with a full model, remove variables one-by-one. Does not work when $p > n$

Mixed selection: combination of forward and backward selection

Q3: Accuracy of the model

Residual standard error (RSE)

$$\mathrm{RSE} = \sqrt{\frac{1}{n-p-1}\mathrm{SSR}} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

an absolute measure of lack of fit

R squared (R^2)

$$\mathrm{R}^2 = 1 - \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \in [0,1]$$

proportion of the variability in Y that can be explained using X

R^2 will **always increase** as more predictors are added to the model!

Both **RSE** and **R squared** favor **flexible** methods , which may **overfit** the data!!

In summary

Linear Regression: simple and multiple

Coefficient estimates

Assessing the accuracy of the coefficient estimates

Next…

Other practical considerations in regression

(Hopefully) new perspectives on regression

Practical considerations in regression

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon$$

Qualitative predictors

Extensions of the linear structures in $(X_1, \ldots, X_p)$

Linear regression diagnostics

Qualitative predictors

In regression setting, Y is quantitative

But some of the predictors $X_1, \ldots, X_p$ can be qualitative

Qualitative predictors: categorical predictors or factor variables

Example: investigate the relationship between credit card balance and the gender of the card holder

Predictor gender takes 2 levels: female or male

Qualitative predictors with only 2 levels

Indicator **(or** **dummy)** **variable that takes on two numeric values**

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

This newly constructed variable x_i can be used as a predictor

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is male} \end{cases}$$

β_1: average difference in credit card balance between female and males

Qualitative predictors with only 2 levels

The coding of the indicator (or dummy) variable is not unique

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

or

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is male} \\ 0 & \text{if } i\text{th person is female} \end{cases}$$

or

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is male} \\ -1 & \text{if } i\text{th person is female} \end{cases}$$

The model can still be written as $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

But the interpretation of β_0 **and** β_1 **depend on the coding of** x_i

Qualitative predictors with more than 2 levels

A qualitative predictor with m levels need $m-1$ dummy variables

e.g., ethnicity variable has three levels: African American, Asian, Caucasian

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is Asian} \\ 0 & \text{if } i\text{th person is not Asian} \end{cases} \qquad x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is Caucasian} \\ 0 & \text{if } i\text{th person is not Caucasian} \end{cases}$$

The level of African American is the baseline level, i.e., no dummy variable needed

Constructed variable x_{i1}, x_{i2} can be used as predictors

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon \\ &= \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{if } i\text{th person is Asian} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{if } i\text{th person is Caucasian} \\ \beta_0 + \varepsilon_i & \text{if } i\text{th person is African American} \end{cases} \end{aligned}$$

Both the dummy coding and the choice of baseline level are arbitrary

But they change the interpretation of coefficients $\beta_0, \beta_1, \beta_2 \ldots$

Multiple regression

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon$$

Qualitative predictors

Extensions of the linear structures in $(X_1, \ldots, X_p)$

Linear regression diagnostics

Extensions of the linear model

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon$$

Relationship between Y and $X_1, \ldots, X_p$ is additive and linear

Additive: the effect of changes in a predictor X_j on Y is independent of the values of all other predictors

Linear: the change in the response Y due to one unit change of a predictor X_j is constant, regardless of the value of X_j

Removing the additive assumption

One way of removing the additive assumption is to include interaction term

from

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

One-unit increase in X_1 results in β_1-unit increase in Y (holding X_2 fixed)

to

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \boxed{X_1 X_2} + \varepsilon$$

interaction between X_1 and X_2

$$= \beta_0 + \boxed{(\beta_1 + \beta_3 X_2)} X_1 + X_2 \beta_2 + \varepsilon$$

One-unit increase in X_1 results in $\beta_1 + \beta_3 X_2$-unit increase in Y

the effect of X_1 on Y is no longer constant: adjusting X_2 will change the impact of X_1 on Y

Interaction: a data example

$$\textbf{sales} = \beta_0 + \beta_1 \times \textbf{TV} + \beta_2 \times \textbf{radio} + \beta_3 \times \textbf{TV} \times \textbf{radio} + \varepsilon$$
$$= \beta_0 + (\beta_1 + \beta_3 \times \textbf{radio}) \times \textbf{TV} + \beta_2 \times \textbf{radio} + \varepsilon$$

		Coefficient	Std. error	t-statistic	p-value
$\hat{\beta}_0$	Intercept	6.7502	0.248	27.23	< 0.0001
$\hat{\beta}_1$	TV	0.0191	0.002	12.70	< 0.0001
$\hat{\beta}_2$	radio	0.0289	0.009	3.24	0.0014
$\hat{\beta}_3$	TV×radio	0.0011	0.000	20.73	< 0.0001

Interpretation: an **increase** in **TV advertising** of **$1,000** is associated with $(\hat{\beta}_1 + \hat{\beta}_3 \times \textbf{radio}) \times 1000 = 19.1 + 1.1 \times \textbf{radio}$ dollars **increase** in sales

Practice: what is the effect of an increase in radio advertising of $1,000 on sales?

Extensions of the linear model

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon$$

Relationship between Y and $X_1, \ldots, X_p$ is additive and linear

Additive: the effect of changes in a predictor X_j on Y is independent of the values of all other predictors

Interaction terms!

Linear: the change in the response Y due to one unit change of a predictor X_j is constant, regardless of the value of X_j

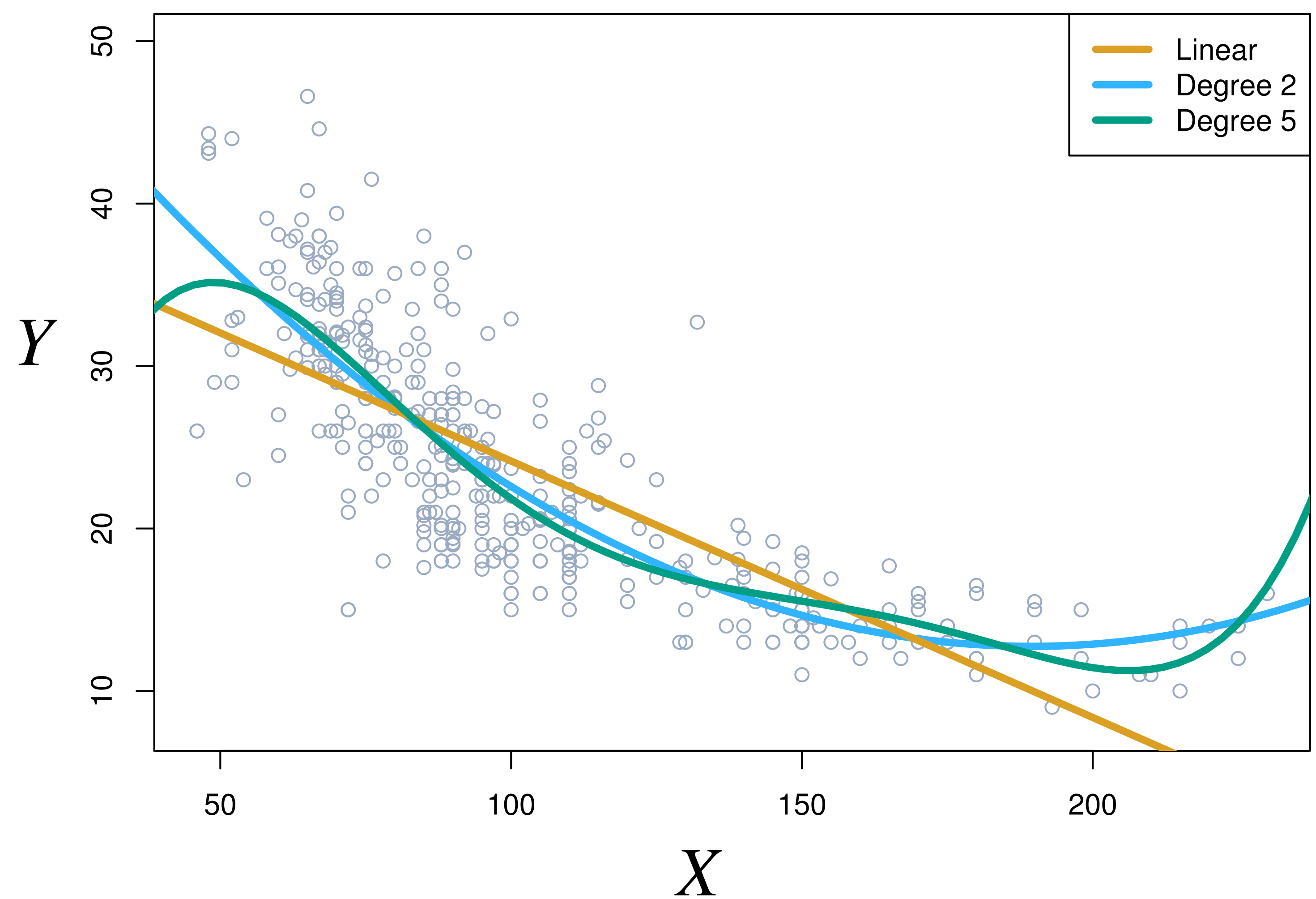
polynomial regression

Non-linear relationships

Polynomial regression of Y on X

degree

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_d X^{\boxed{d}} + \varepsilon$$



Non-linear relationships

Polynomial regression of Y on X

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_d X^d + \varepsilon$$

Y **is no longer linear in** X

But this is still a linear model!!!

Simply let $Z_k = X^k$**...**

$$Y = \beta_0 + \beta_1 X + \beta_2 Z_2 + \ldots + \beta_d Z_d + \varepsilon$$

Y **is still linear in** $X, Z_2, \ldots, Z_d$

Practical considerations in regression

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon$$

Qualitative predictors

Extensions of the linear structures in $(X_1, \ldots, X_p)$

Linear regression diagnostics

Linear regression diagnostics

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon$$

1. **Non-linear relationship between response and predictors**
2. **Correlation of error terms**
3. **Non-constant variance of error terms**
4. **Outliers**
5. **High-leverage points**
6. **Collinearity**

Non-linearity of the data

We assume…

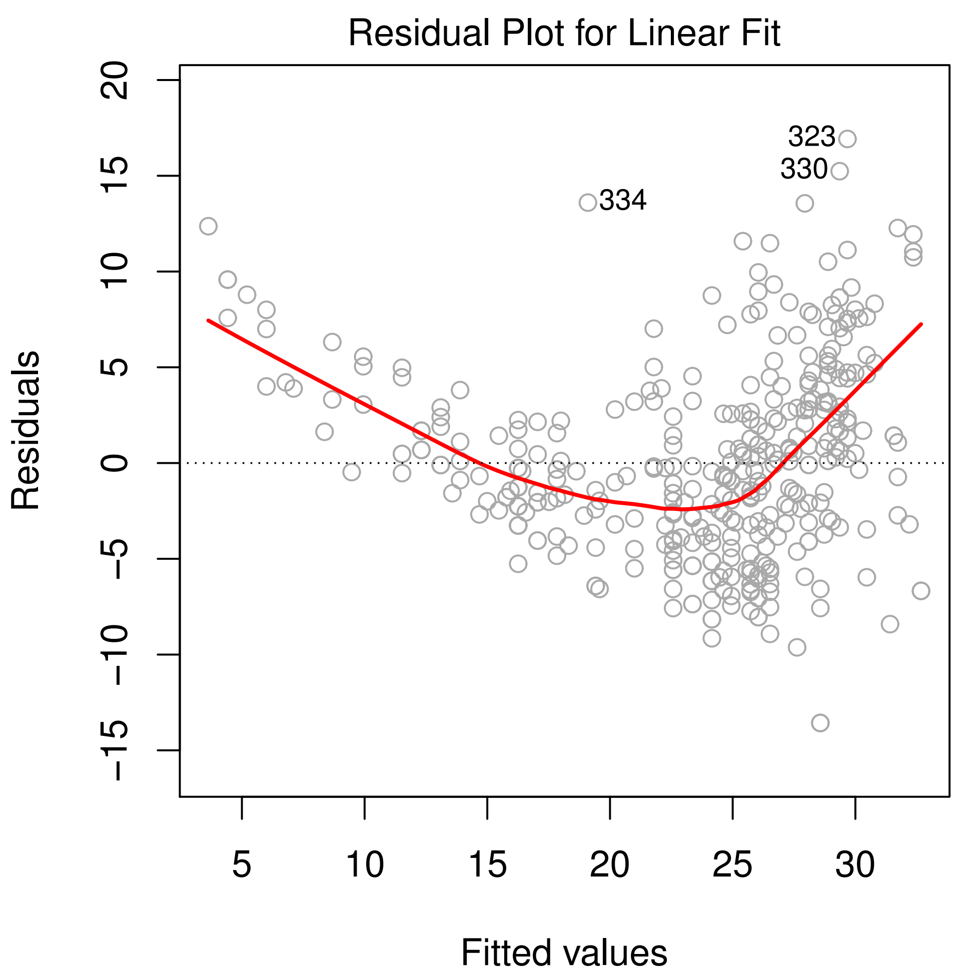
$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon$$

but this is not necessarily the true relationship between Y and $X_1, \ldots, X_p$

How can we tell if we made the wrong linear assumption?

Residual plot

plot the residual $y_i - \hat{y}_i$ v.s the fitted value (prediction) $\hat{y}_i$



If our linear assumption is correct, then the residual plot will NOT show discernible pattern

Correlation of error terms

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i$$

We assume that $\varepsilon_1, \ldots, \varepsilon_n$ are uncorrelated

If they are correlated, then the estimated standard errors will not be accurate

Inaccurate confidence interval or hypothesis testing results

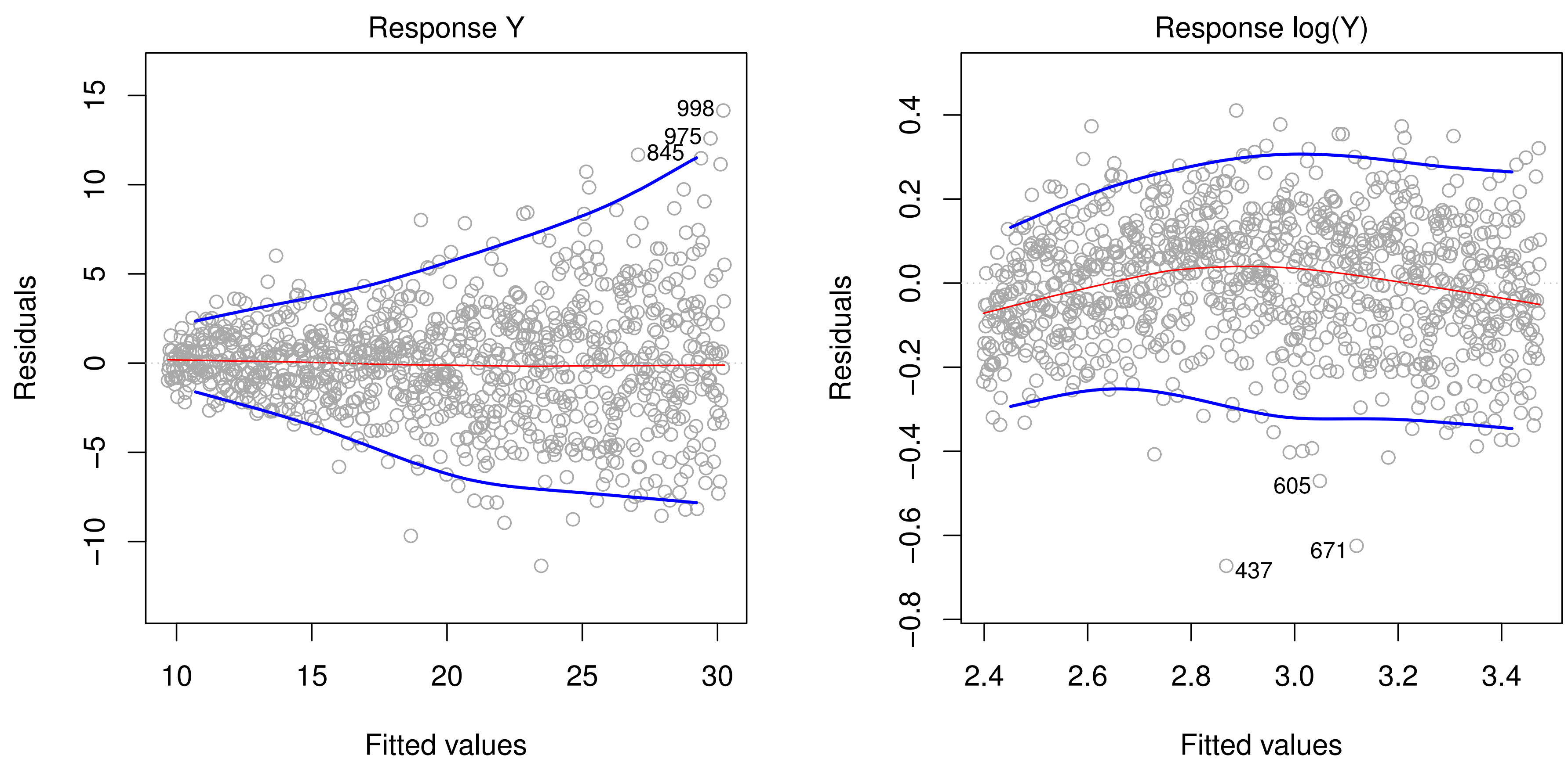
Non-constant variance of error terms

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i$$

We also assume that $\mathrm{Var}(\varepsilon_i) = \sigma^2$ for all $i = 1, \ldots, n$

Heteroscedasticity: Non-constant variances in the errors

Inaccurate confidence interval or hypothesis testing results



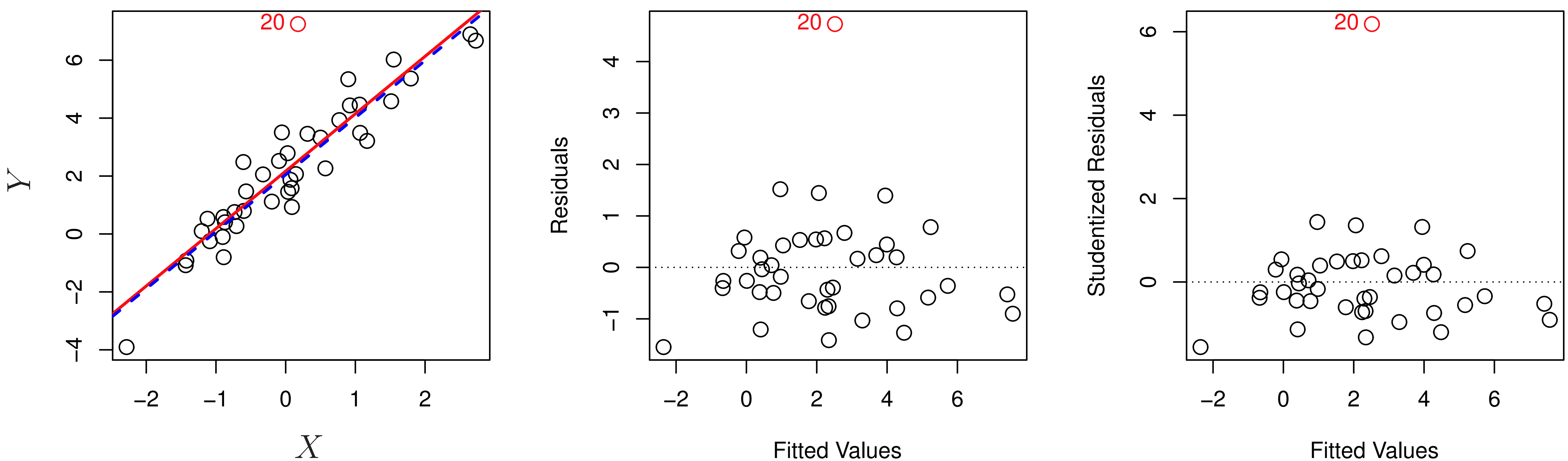
Outliers

Outlier: a data point for which y_i is far from $\hat{y}_i$ given by the model

e.g., incorrect recording of an observation during data collection

How to find outliers?

Calculate the studentized residuals



High leverage points

High leverage points: a data point for which $x_{i1}, \ldots, x_{ip}$ have unusual values

How to find high leverage points?

Calculate the leverage statistics (for simple linear regression)

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n} (x_k - \bar{x})^2}$$

Collinearity

Collinearity: two or more predictors are closely related to one another

collinear

If two predictors tend to increase or decrease together, it can be difficult to determine how each one is associated with the response

The variance of the estimates increase

How to detect collinearity?

Approach 1: look at correlation matrix of $X_1, \ldots, X_p$

Approach 2: compute the variance inflation factor

How to handle collinearity between, say, X_1 and X_2?

Approach 1: drop one of X_1, X_2 in regression model

Approach 2: combine X_1 and X_2 (hard to interpret)

Next…

Other practical considerations in regression

New perspectives on regression

Linear regression vs K-NN regression

in the general regression setting $Y = f(X) + \varepsilon$

Parametric approach	Non-parametric approach
Assume that $f(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$	$f(X)$ can have any function form
No need to tune the model	Tuning parameter: K
Performs well when the true $f(X)$ is close to linear	Much more general-purpose
Interpretability, statistical inference…	Not very interpretable
Can be extended to work when p is very large ridge regression, lasso …	Curse of Dimensionality

Bias-Variance tradeoff in linear regression

Assume that $Y = f(X) + \varepsilon = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$

$$\mathrm{E}\left[\left(y_0 - \hat{f}(\mathbf{x}_0)\right)^2\right] = \mathrm{Var}(\hat{f}(\mathbf{x}_0)) + \left[\mathrm{Bias}(\hat{f}(\mathbf{x}_0))\right]^2 + \mathrm{Var}(\varepsilon),$$

For $\hat{f} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}_{01} + \ldots + \hat{\beta}_p \mathbf{x}_{0p}$**, where** $\hat{\beta}_0, \ldots, \hat{\beta}_p$ **are least-squares estimates**

Property 1: Unbiased, i.e., $\mathrm{Bias}(\hat{f}(\mathbf{x}_0)) = \mathrm{E}[\hat{f}(\mathbf{x}_0)] - f(\mathbf{x}_0) = 0$

Property 2: Least-squares has the **smallest** expected test error among all **unbiased linear** estimates (**Gauss-Markov Theorem**)

Modern regression methods can **outperform** least-squares in terms of expected test MSE, by **having small bias** but **having much smaller variance**

In summary

Practical considerations in regression

Qualitative predictors

Extensions of the linear structures in $(X_1, \ldots, X_p)$

Linear regression diagnostics

1. Non-linear relationship between response and predictors
2. Correlation of error terms
3. Non-constant variance of error terms
4. Outliers
5. High-leverage points
6. Collinearity

New perspectives on regression

Compare linear regression with K-NN regression

Bias-variance tradeoff of linear regression

Next…

Linear Classification method: logistic regression

Quiz 1 on Friday!