6.7900: Machine Learning
Lecture 4

Lecture start: Tues/Thurs 2:35pm
Who's speaking today? Prof. Tamara Broderick
Course website: gradml.mit.edu
Questions? Ask here or on piazza.com/mit/fall2024/67900/
Materials: Slides, video, etc linked from gradml.mit.edu after the lecture (but there is no livestream)

Last Times

I. Empirical risk minimization (ERM)
II. Maximum likelihood estimate (MLE)
III. No features

Today's Plan

I. MLE & ERM for supervised learning
II. Linear regression: Why and how
III. Some challenges

Why linear regression?

Why linear regression?

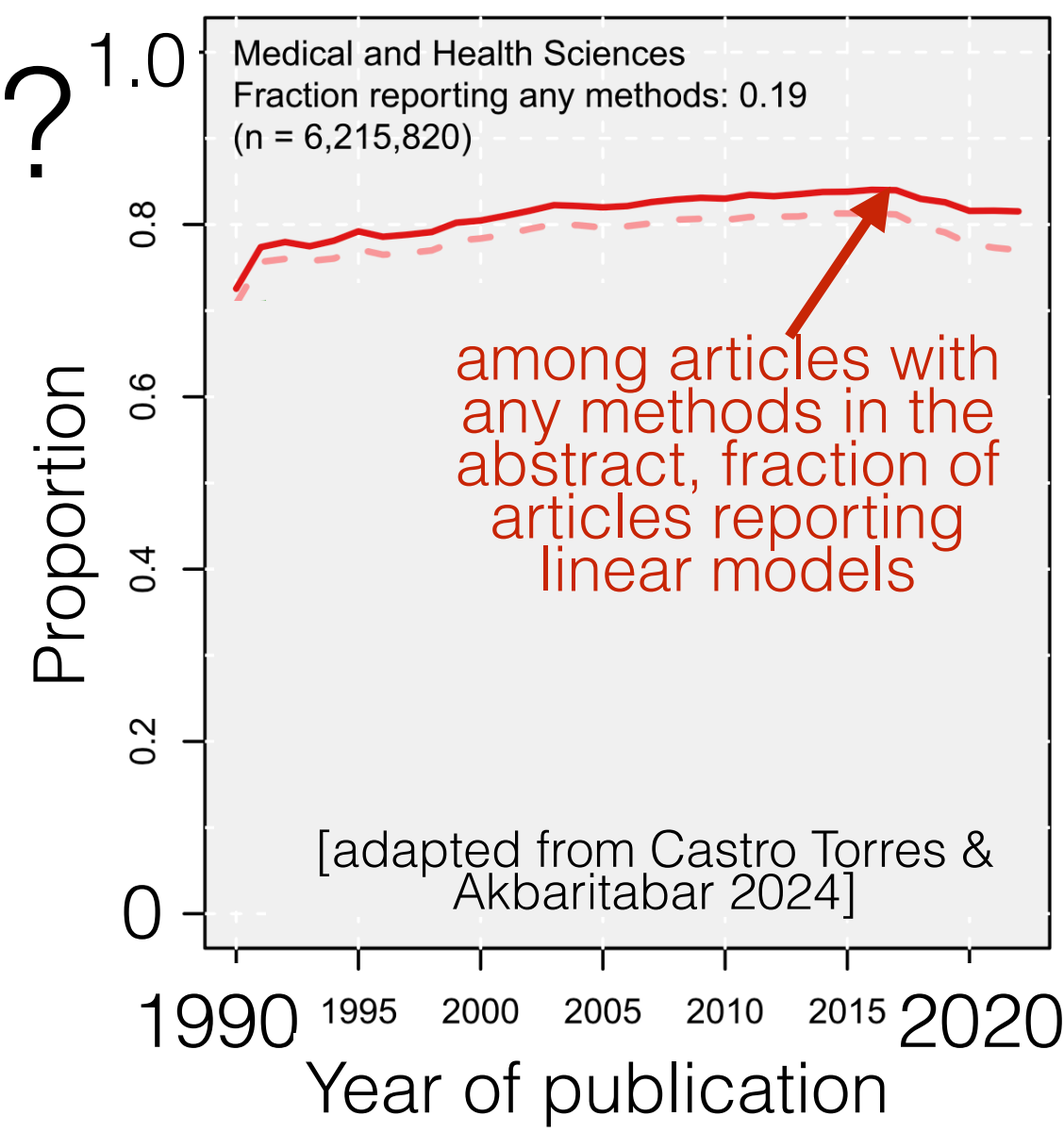
- Do modern, sophisticated robots make hammers obsolete?

Why linear regression?

- Do modern, sophisticated robots make hammers obsolete?
- Widely used

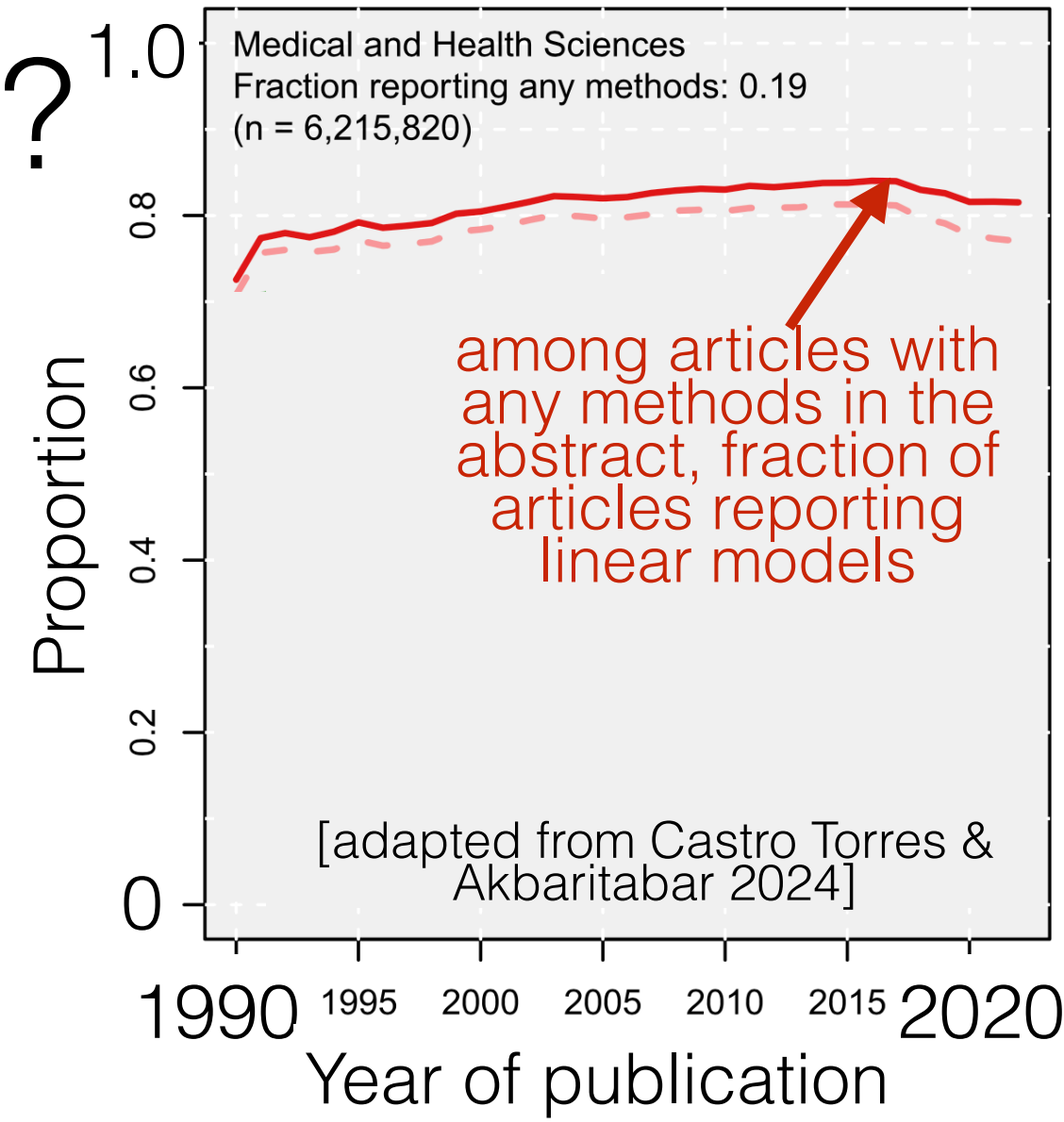
Why linear regression?

- Do modern, sophisticated robots make hammers obsolete?
- Widely used



Why linear regression?

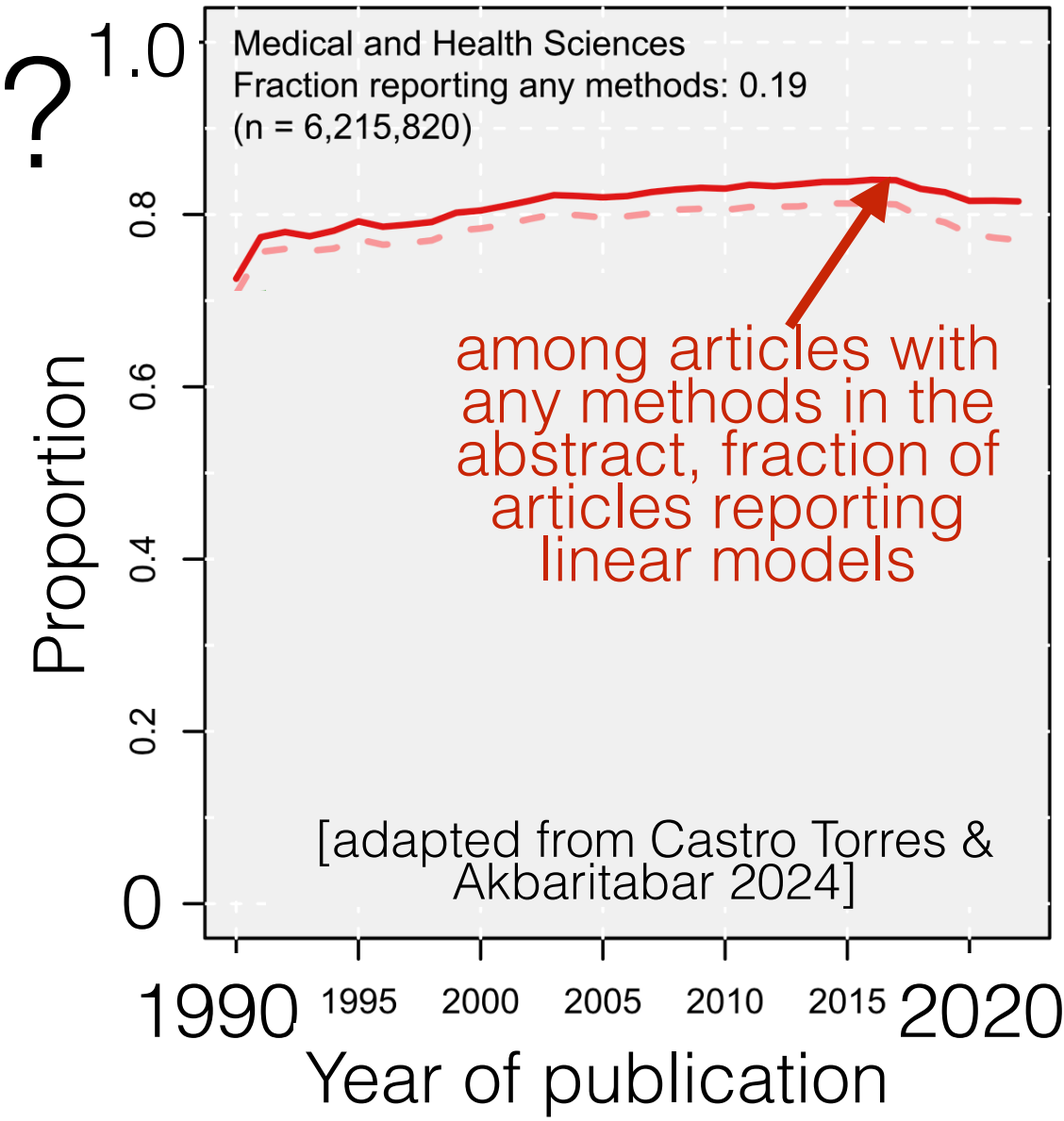
- Do modern, sophisticated robots make hammers obsolete?
- Widely used
- Speed, interpretability



Why linear regression?

- Do modern, sophisticated robots make hammers obsolete?
- Widely used
- Speed, interpretability
- Theoretical insight, including about deep learning

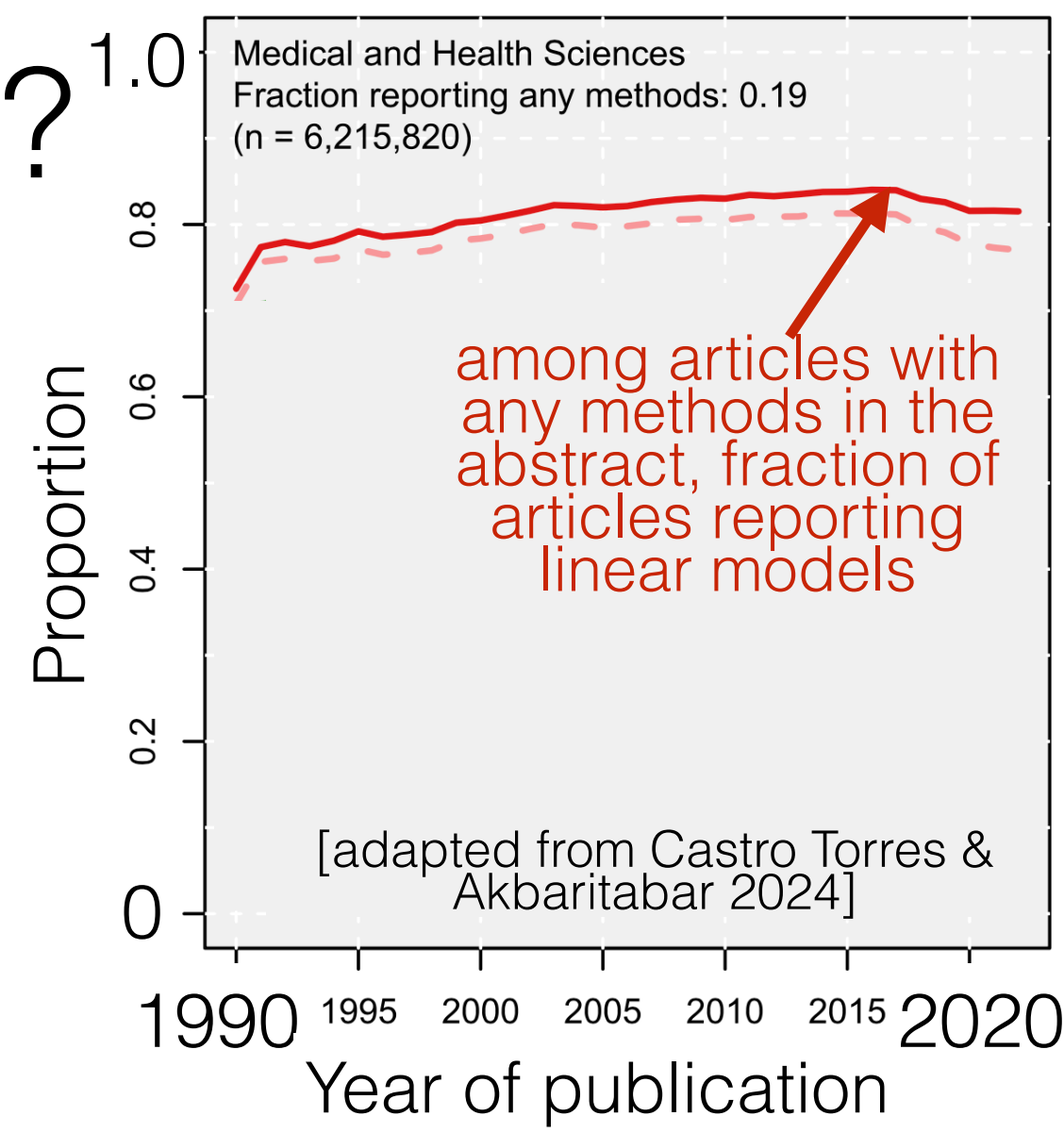
[e.g. Bartlett+ 2020, Benign overfitting in linear regression]



Why linear regression?

- Do modern, sophisticated robots make hammers obsolete?
- Widely used
- Speed, interpretability
- Theoretical insight, including about deep learning
- Often a cog in modern procedures

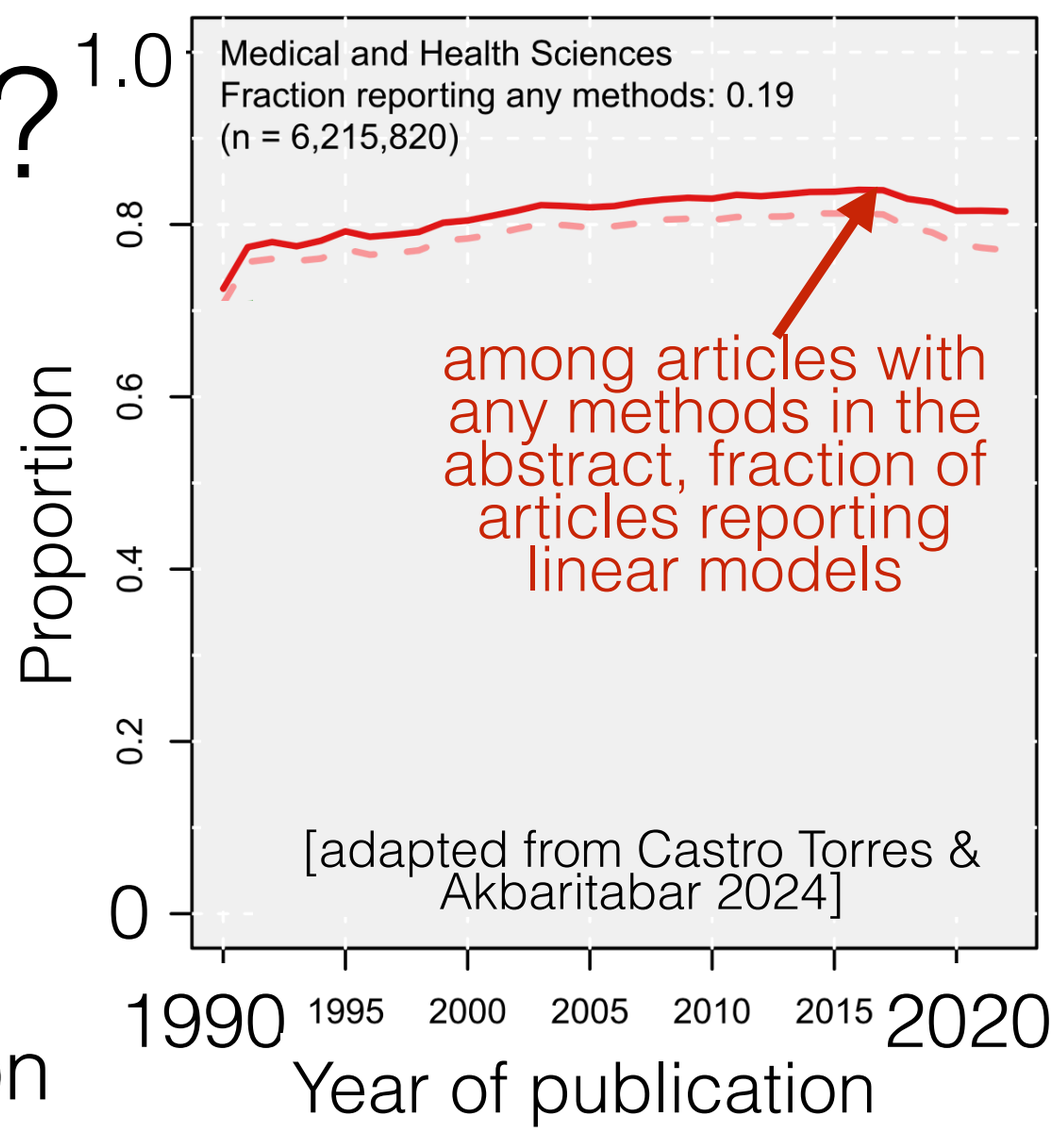
[e.g. Bartlett+ 2020, Benign overfitting in linear regression]



Why linear regression?

- Do modern, sophisticated robots make hammers obsolete?
- Widely used
- Speed, interpretability
- Theoretical insight, including about deep learning
- Often a cog in modern procedures
- Importance of baselines and ablation

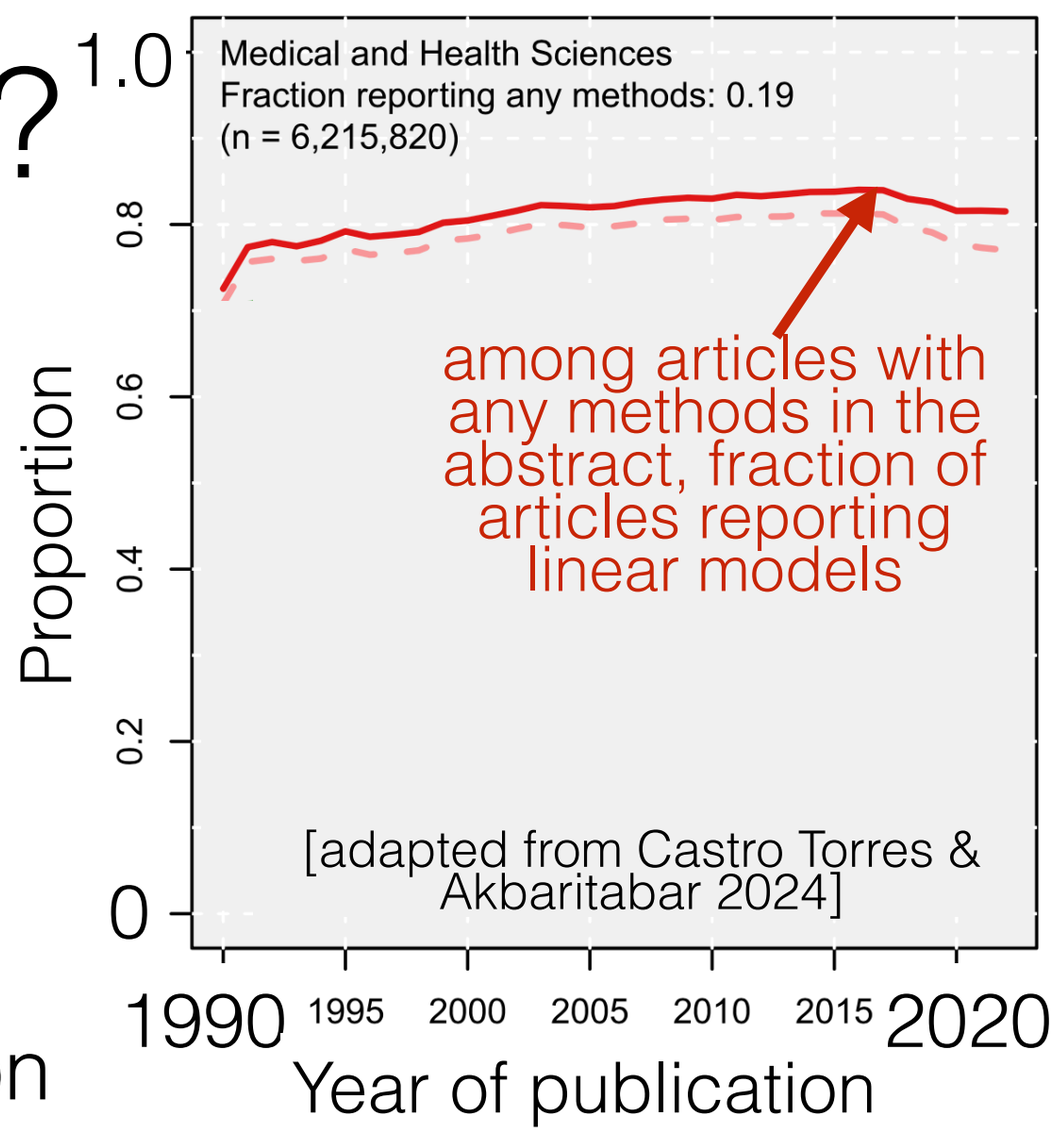
[e.g. Bartlett+ 2020, Benign overfitting in linear regression]



Why linear regression?

- Do modern, sophisticated robots make hammers obsolete?
- Widely used
- Speed, interpretability
- Theoretical insight, including about deep learning
- Often a cog in modern procedures
- Importance of baselines and ablation
  - Lipton, Steinhardt 2019, Troubling Trends in Machine Learning Scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research

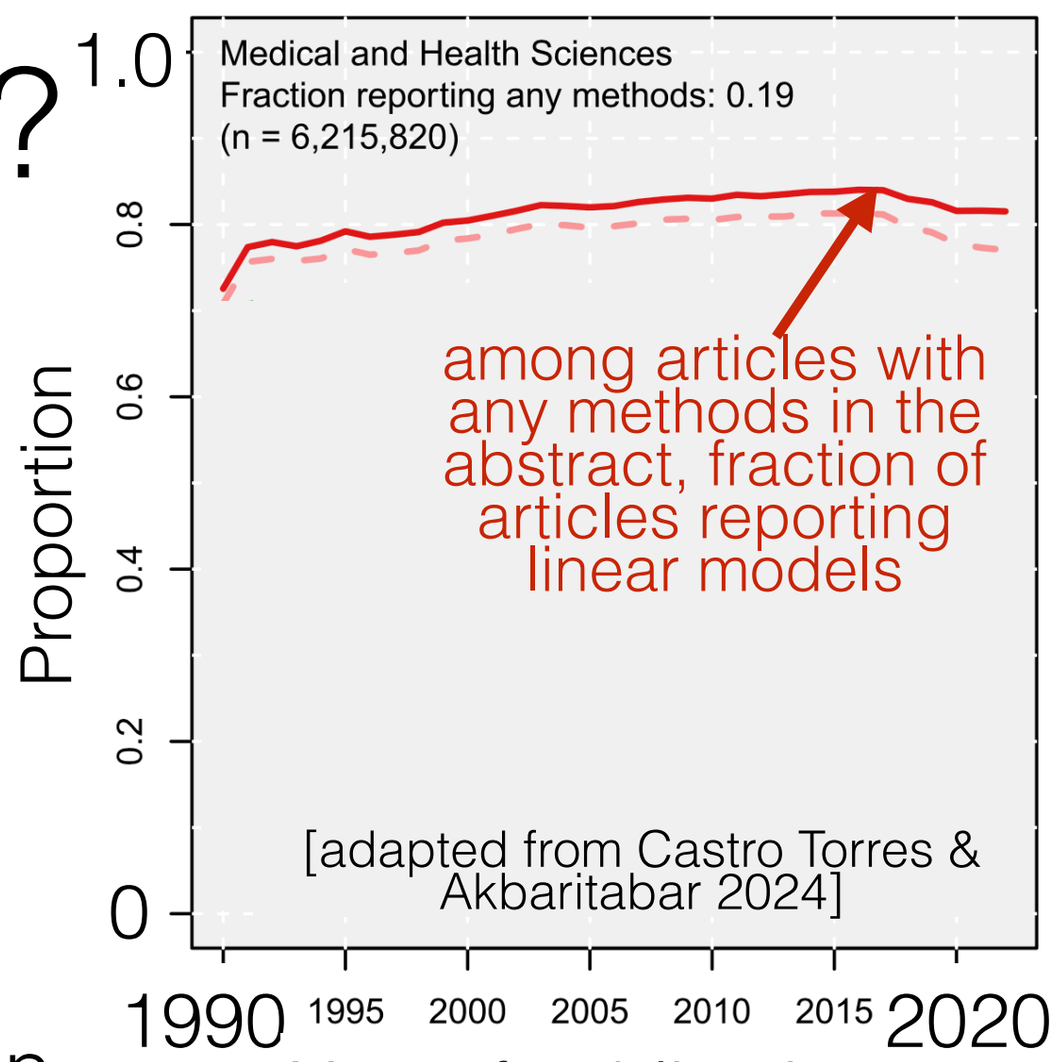
[e.g. Bartlett+ 2020, Benign overfitting in linear regression]



Why linear regression?

- Do modern, sophisticated robots make hammers obsolete?
- Widely used
- Speed, interpretability
- Theoretical insight, including about deep learning
- Often a cog in modern procedures
- Importance of baselines and ablation
  - Lipton, Steinhardt 2019, Troubling Trends in Machine Learning Scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research
  - Salganik+ 2020: “100s of researchers attempted to predict 6 life outcomes, such as a child’s GPA and whether a family would be evicted[….T]hey drew on a vast dataset that was painstakingly collected […] over 15y. [N]o one made very accurate predictions.”

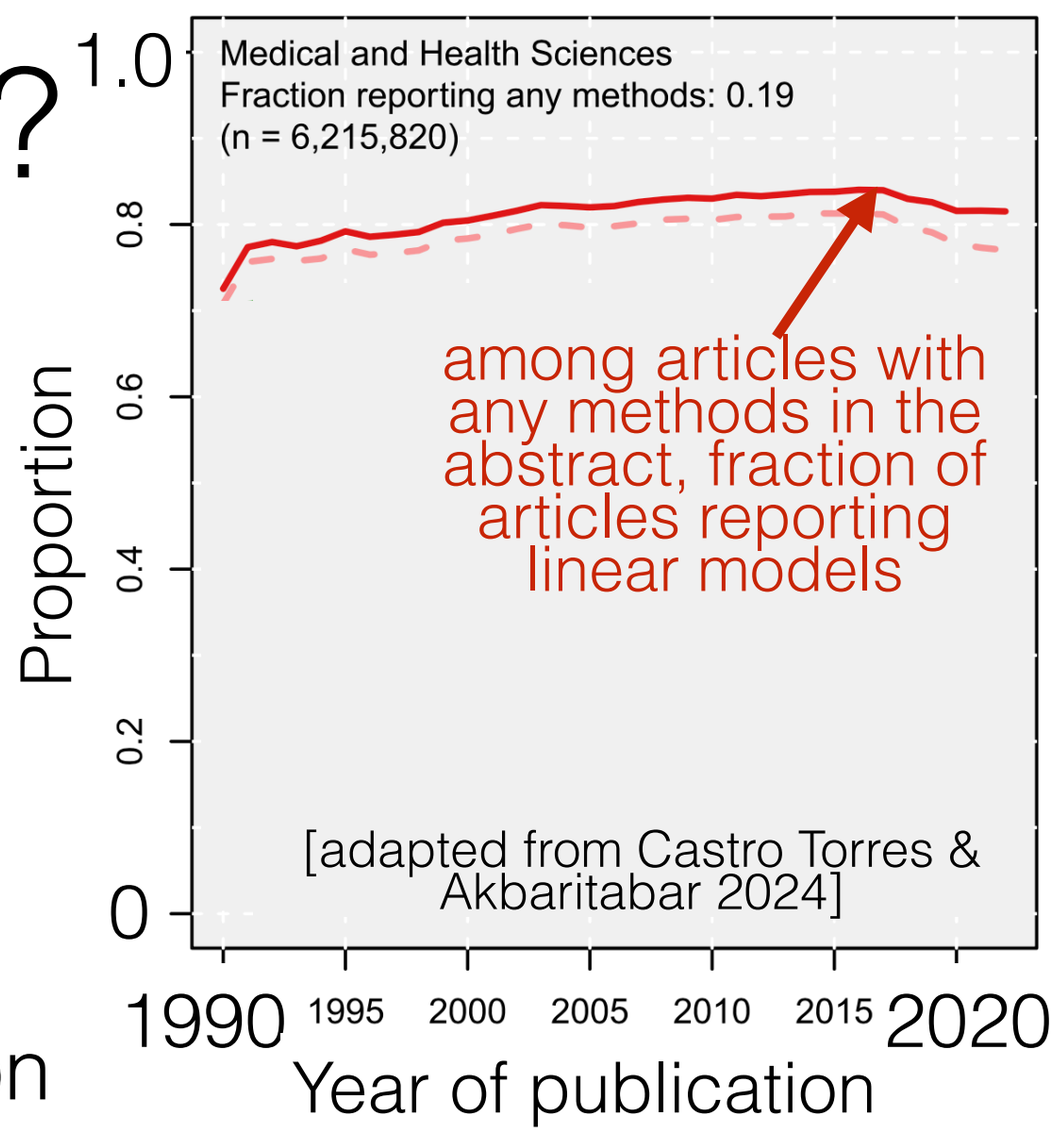
[e.g. Bartlett+ 2020, Benign overfitting in linear regression]



Why linear regression?

- Do modern, sophisticated robots make hammers obsolete?
- Widely used
- Speed, interpretability
- Theoretical insight, including about deep learning
- Often a cog in modern procedures
- Importance of baselines and ablation
  - Lipton, Steinhardt 2019, Troubling Trends in Machine Learning Scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research
  - Salganik+ 2020: **“100s of researchers attempted to predict 6 life outcomes, such as a child’s GPA and whether a family would be evicted[….T]hey drew on a vast dataset that was painstakingly collected […] over 15y. [N]o one made very accurate predictions.”**
  - Chen et al 2016, A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task

[e.g. Bartlett+ 2020, Benign overfitting in linear regression]



Motivating example

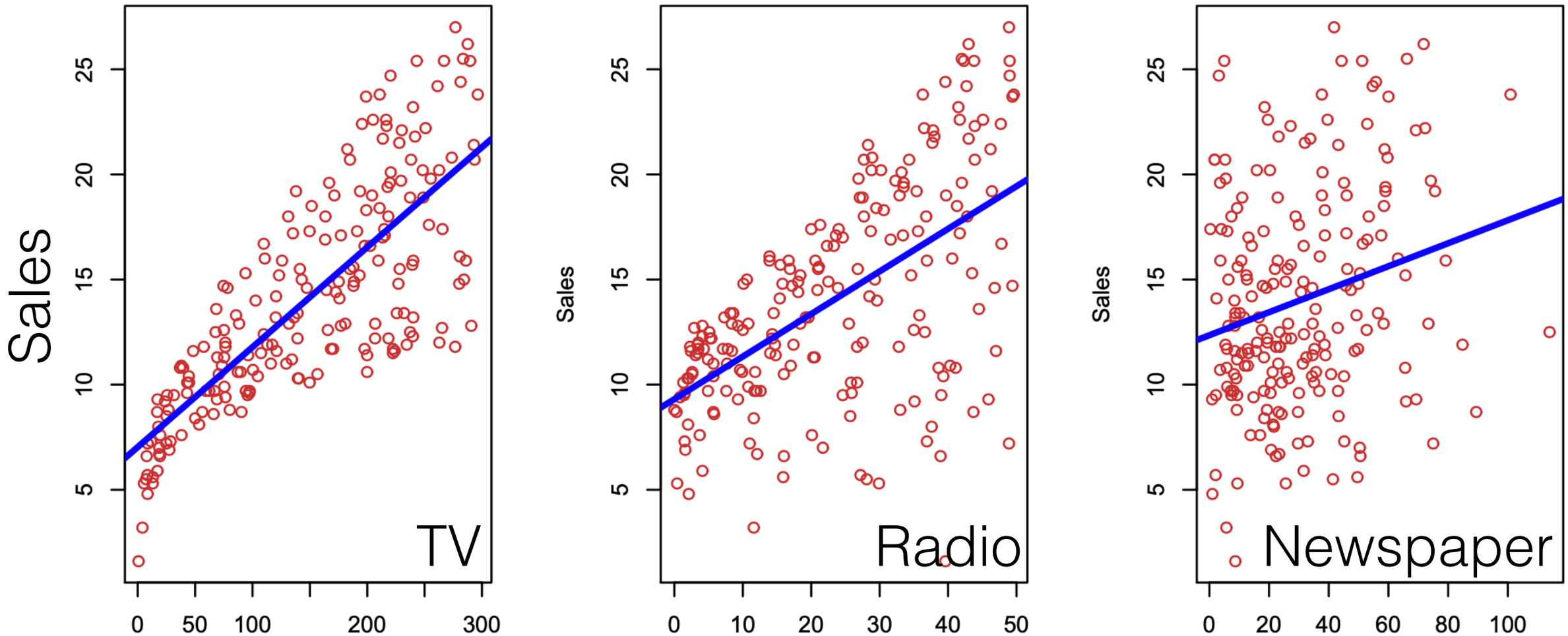
Motivating example

- Advertising data: “sales, in thousands of units, as a function of TV, radio, and newspaper budgets, in thousands of dollars, for 200 different markets”

Motivating example

- Advertising data: “sales, in thousands of units, as a function of TV, radio, and newspaper budgets, in thousands of dollars, for 200 different markets”

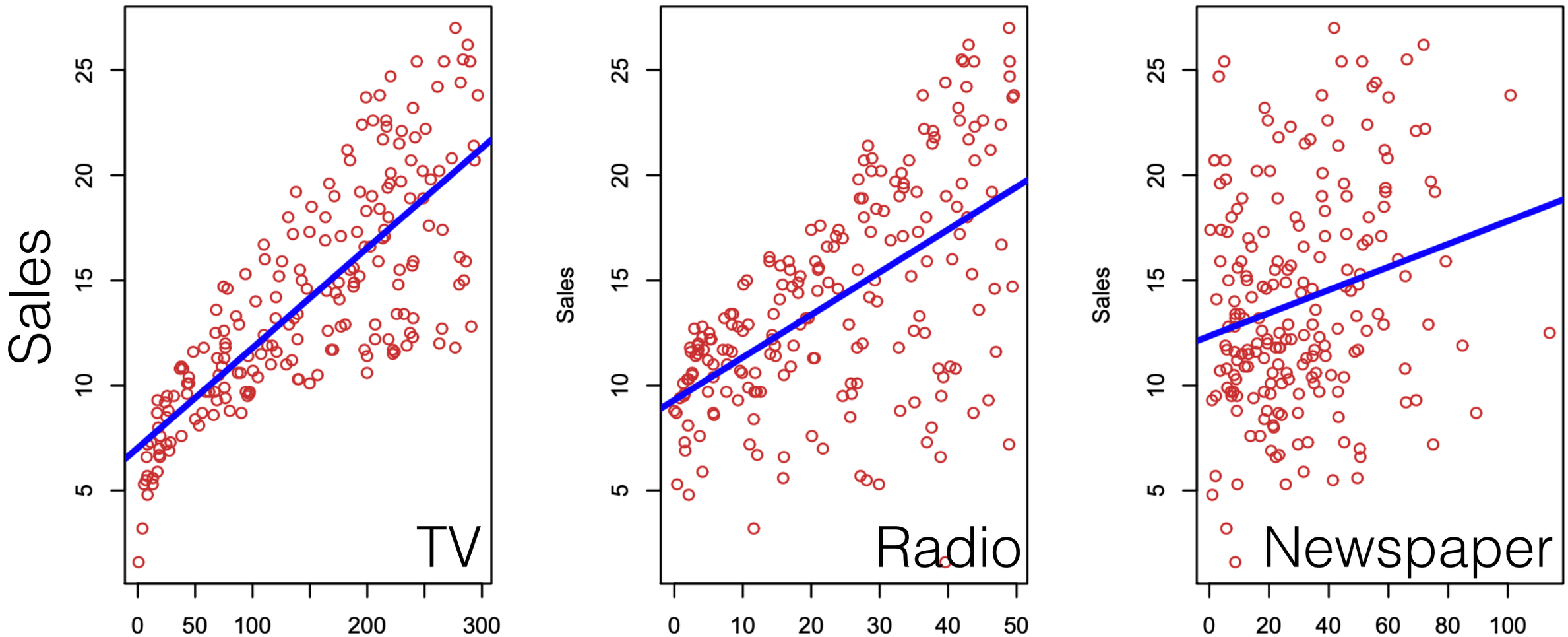
[An Intro to Statistical Learning 2023, Fig 2.1]



Motivating example

- Advertising data: “sales, in thousands of units, as a function of TV, radio, and newspaper budgets, in thousands of dollars, for 200 different markets”

[An Intro to Statistical Learning 2023, Fig 2.1]

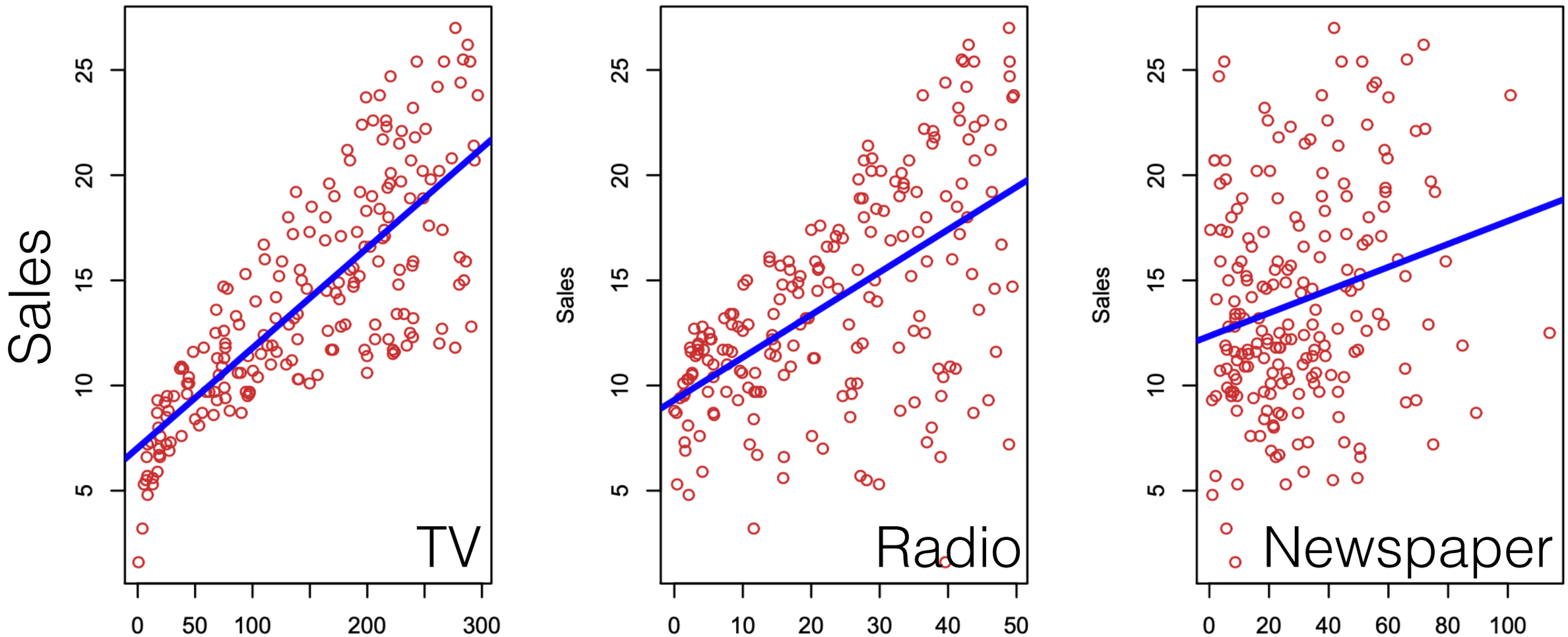


- Data point: label is sales, in thousands of units: $y^{(n)} \in \mathbb{R}_+$
  - Features: $x^{(n)} = [x_1^{(n)}, x_2^{(n)}, x_3^{(n)}]^\top \in \mathbb{R}_+^3$

Motivating example

- Advertising data: “sales, in thousands of units, as a function of TV, radio, and newspaper budgets, in thousands of dollars, for 200 different markets”

[An Intro to Statistical Learning 2023, Fig 2.1]



- Data point: label is sales, in thousands of units: $y^{(n)} \in \mathbb{R}_+$
  - Features: $x^{(n)} = [x_1^{(n)}, x_2^{(n)}, x_3^{(n)}]^\top \in \mathbb{R}_+^3$
- Want to predict sales under different ad budgets, allocate ad budget. Separate regressions not same as whole.

Choosing a likelihood

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
  - Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}$

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
  - Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}, \epsilon^{(n)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
  - Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}, \epsilon^{(n)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

1x1 Dx1

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
  - Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}, \epsilon^{(n)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

1x1 1xD Dx1

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
  - Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}, \epsilon^{(n)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
    1x1 1xD Dx1 1x1

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
  - Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}, \epsilon^{(n)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
    1x1 1xD Dx1 1x1 1x1

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
- Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}, \epsilon^{(n)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

1x1 1xD Dx1 1x1 1x1

"All models are wrong" What about this one?

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
- Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}, \epsilon^{(n)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

1x1 1xD Dx1 1x1 1x1

homoskedastic noise

"All models are wrong" What about this one?

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
- Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}, \epsilon^{(n)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
  - Can handle an intercept term by taking: $\forall n, x_1^{(n)} = 1$

homoskedastic noise

"All models are wrong" What about this one?

1x1 1xD Dx1 1x1 1x1

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
- Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}, \epsilon^{(n)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
  - Can handle an intercept term by taking: $\forall n, x_1^{(n)} = 1$
  - Equivalently:

homoskedastic noise

"All models are wrong"

1x1 1xD Dx1 1x1 1x1

What about this one?

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
- Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}, \epsilon^{(n)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
  - Can handle an intercept term by taking: $\forall n, x_1^{(n)} = 1$
  - Equivalently: $y^{(n)}|x^{(n)} \overset{indep}{\sim} \mathcal{N}(\theta^\top x^{(n)}, \sigma^2)$

homoskedastic noise

"All models are wrong"

1x1 1xD Dx1 1x1 1x1

What about this one?

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
- Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}, \epsilon^{(n)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
  - 1x1 1xD Dx1 1x1 1x1
  - Can handle an intercept term by taking: $\forall n, x_1^{(n)} = 1$
  - Equivalently: $y^{(n)}|x^{(n)} \overset{indep}{\sim} \mathcal{N}(\theta^\top x^{(n)}, \sigma^2)$

$$p(y|x, \theta, \sigma^2) = \mathcal{N}(y|\theta^\top x, \sigma^2)$$

homoskedastic noise

"All models are wrong" What about this one?

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
- Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}, \epsilon^{(n)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ (1x1 1xD Dx1 1x1; 1x1)
  - Can handle an intercept term by taking: $\forall n, x_1^{(n)} = 1$
  - Equivalently: $y^{(n)}|x^{(n)} \overset{indep}{\sim} \mathcal{N}(\theta^\top x^{(n)}, \sigma^2)$

    $$p(y|x, \theta, \sigma^2) = \mathcal{N}(y|\theta^\top x, \sigma^2)$$

  - Taking the MLE approach we've described so far:

homoskedastic noise

"All models are wrong" What about this one?

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
- Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}, \epsilon^{(n)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$
  (1x1 = $y^{(n)}$, 1xD = $\theta^\top$, Dx1 = $x^{(n)}$, 1x1 = $\epsilon^{(n)}$, 1x1 = $\mathcal{N}(0, \sigma^2)$)
  - Can handle an intercept term by taking: $\forall n, x_1^{(n)} = 1$
  - Equivalently: $y^{(n)}|x^{(n)} \overset{indep}{\sim} \mathcal{N}(\theta^\top x^{(n)}, \sigma^2)$

$$p(y|x, \theta, \sigma^2) = \mathcal{N}(y|\theta^\top x, \sigma^2)$$

  - Taking the MLE approach we've described so far:

fit params $= \arg\max p(\{y^{(n)}\}_1^N | \{x^{(n)}\}_1^N,$ parameters?)

parameters

homoskedastic noise

"All models are wrong" What about this one?

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
  - Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}, \epsilon^{(n)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ (1x1 1xD Dx1 1x1; 1x1)
    - Can handle an intercept term by taking: $\forall n, x_1^{(n)} = 1$
    - Equivalently: $y^{(n)}|x^{(n)} \overset{indep}{\sim} \mathcal{N}(\theta^\top x^{(n)}, \sigma^2)$

$$p(y|x, \theta, \sigma^2) = \mathcal{N}(y|\theta^\top x, \sigma^2)$$

    - Taking the MLE approach we've described so far:

$$\hat{\theta}, \hat{\sigma}^2 = \arg\max_{\theta, \sigma^2} p(\{y^{(n)}\}_1^N | \{x^{(n)}\}_1^N, \theta, \sigma^2)$$

homoskedastic noise

"All models are wrong" What about this one?

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
  - Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}, \epsilon^{(n)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ (1x1 1xD Dx1 1x1; 1x1)
    - Can handle an intercept term by taking: $\forall n, x_1^{(n)} = 1$
    - Equivalently: $y^{(n)}|x^{(n)} \overset{indep}{\sim} \mathcal{N}(\theta^\top x^{(n)}, \sigma^2)$

$$p(y|x, \theta, \sigma^2) = \mathcal{N}(y|\theta^\top x, \sigma^2)$$

    - Taking the MLE approach we've described so far:

$$\hat{\theta}, \hat{\sigma}^2 = \arg\max_{\theta, \sigma^2} p(\{y^{(n)}\}_1^N | \{x^{(n)}\}_1^N, \theta, \sigma^2) \quad \text{\& use} \quad p(y|x, \hat{\theta}, \hat{\sigma}^2)$$

homoskedastic noise

"All models are wrong" What about this one?

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
  - Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}, \epsilon^{(n)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ (1x1 1xD Dx1 1x1 1x1)
    - Can handle an intercept term by taking: $\forall n, x_1^{(n)} = 1$
    - Equivalently: $y^{(n)}|x^{(n)} \overset{indep}{\sim} \mathcal{N}(\theta^\top x^{(n)}, \sigma^2)$

$$p(y|x, \theta, \sigma^2) = \mathcal{N}(y|\theta^\top x, \sigma^2)$$

- Taking the MLE approach we've described so far:

$$\hat{\theta}, \hat{\sigma}^2 = \arg\max_{\theta, \sigma^2} p(\{y^{(n)}\}_1^N | \{x^{(n)}\}_1^N, \theta, \sigma^2) \quad \text{\& use} \quad p(y|x, \hat{\theta}, \hat{\sigma}^2)$$

- So here predict $h(x) = \mathbb{E}[Y|X = x] =$?

homoskedastic noise

"All models are wrong" What about this one?

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
  - Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}, \epsilon^{(n)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ (1x1 1xD Dx1 1x1; 1x1)
    - Can handle an intercept term by taking: $\forall n, x_1^{(n)} = 1$
    - Equivalently: $y^{(n)}|x^{(n)} \overset{indep}{\sim} \mathcal{N}(\theta^\top x^{(n)}, \sigma^2)$

$$p(y|x, \theta, \sigma^2) = \mathcal{N}(y|\theta^\top x, \sigma^2)$$

    - Taking the MLE approach we've described so far:

$$\hat{\theta}, \hat{\sigma}^2 = \arg\max_{\theta, \sigma^2} p(\{y^{(n)}\}_1^N | \{x^{(n)}\}_1^N, \theta, \sigma^2)$$ & use $p(y|x, \hat{\theta}, \hat{\sigma}^2)$

      - So here predict $h(x) = \mathbb{E}[Y|X = x] = \hat{\theta}^\top x$

homoskedastic noise

"All models are wrong" What about this one?

Choosing a likelihood

Recall proposition from Lecture 1

- **Proposition**. Consider regression with $\mathcal{X} = \mathbb{R}^D, \mathcal{Y} = \mathbb{R}$
  - Assume *X*,*Y* density *p*(*x*,*y*) & square loss $L(a, g) = (a - g)^2$
  - (Assume all needed integrals exist.) Then this decision rule minimizes the risk of a new point: $h(x) = \mathbb{E}[Y|X = x]$
- One approach: we can make a model for $p(y|x)$ and then use the maximum likelihood estimate of the parameter
  - Due to the form of *h*(*x*), it seems like we don't need to worry about a model for *p*(*x*)
  - Likelihood: $y^{(n)} = \theta^\top x^{(n)} + \epsilon^{(n)}, \epsilon^{(n)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ (1x1 1xD Dx1 1x1; 1x1)
    - Can handle an intercept term by taking: $\forall n, x_1^{(n)} = 1$
    - Equivalently: $y^{(n)}|x^{(n)} \overset{indep}{\sim} \mathcal{N}(\theta^\top x^{(n)}, \sigma^2)$

$$p(y|x, \theta, \sigma^2) = \mathcal{N}(y|\theta^\top x, \sigma^2)$$

- Taking the MLE approach we've described so far:

$$\hat{\theta}, \hat{\sigma}^2 = \arg\max_{\theta, \sigma^2} p(\{y^{(n)}\}_1^N | \{x^{(n)}\}_1^N, \theta, \sigma^2)$$ & use $p(y|x, \hat{\theta}, \hat{\sigma}^2)$

- So here predict $h(x) = \mathbb{E}[Y|X = x] = \hat{\theta}^\top x$ (no variance param!)

homoskedastic noise

"All models are wrong" What about this one?

Maximizing the likelihood

Maximizing the likelihood

- Likelihood of the training data:

$p(\{y^{(n)}\}_1^N | \{x^{(n)}\}_1^N, \theta, \sigma^2)$

Maximizing the likelihood

- Likelihood of the training data:

$$p(\{y^{(n)}\}_1^N|\{x^{(n)}\}_1^N, \theta, \sigma^2) = \prod_{n=1}^N p(y^{(n)}|x^{(n)}, \theta, \sigma^2)$$

Maximizing the likelihood

- Likelihood of the training data:

$$p(\{y^{(n)}\}_1^N | \{x^{(n)}\}_1^N, \theta, \sigma^2) = \prod_{n=1}^N p(y^{(n)} | x^{(n)}, \theta, \sigma^2)$$

$$= \prod_{n=1}^N \mathcal{N}(y^{(n)} | \theta^\top x^{(n)}, \sigma^2)$$

Maximizing the likelihood

- Likelihood of the training data:

$$p(\{y^{(n)}\}_1^N|\{x^{(n)}\}_1^N,\theta,\sigma^2) = \prod_{n=1}^N p(y^{(n)}|x^{(n)},\theta,\sigma^2)$$

$$= \prod_{n=1}^N \mathcal{N}(y^{(n)}|\theta^\top x^{(n)},\sigma^2)$$

$$= \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-(y^{(n)}-\theta^\top x^{(n)})^2/(2\sigma^2)\right\}$$

Maximizing the likelihood

- Likelihood of the training data:

$$p(\{y^{(n)}\}_1^N|\{x^{(n)}\}_1^N,\theta,\sigma^2) = \prod_{n=1}^N p(y^{(n)}|x^{(n)},\theta,\sigma^2)$$

$$= \prod_{n=1}^N \mathcal{N}(y^{(n)}|\theta^\top x^{(n)},\sigma^2)$$

$$= \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-(y^{(n)}-\theta^\top x^{(n)})^2/(2\sigma^2)\right\}$$

- Equivalent to minimize the negative log likelihood (NLL)

Maximizing the likelihood

- Likelihood of the training data:

$$p(\{y^{(n)}\}_1^N|\{x^{(n)}\}_1^N,\theta,\sigma^2) = \prod_{n=1}^N p(y^{(n)}|x^{(n)},\theta,\sigma^2)$$
$$= \prod_{n=1}^N \mathcal{N}(y^{(n)}|\theta^\top x^{(n)},\sigma^2)$$
$$= \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-(y^{(n)}-\theta^\top x^{(n)})^2/(2\sigma^2)\right\}$$

- Equivalent to minimize the negative log likelihood (NLL)

$$-\log p(\{y^{(n)}\}_1^N|\{x^{(n)}\}_1^N,\theta,\sigma^2)$$

Maximizing the likelihood

- Likelihood of the training data:

$$p(\{y^{(n)}\}_1^N|\{x^{(n)}\}_1^N,\theta,\sigma^2) = \prod_{n=1}^N p(y^{(n)}|x^{(n)},\theta,\sigma^2)$$
$$= \prod_{n=1}^N \mathcal{N}(y^{(n)}|\theta^\top x^{(n)},\sigma^2)$$
$$= \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-(y^{(n)}-\theta^\top x^{(n)})^2/(2\sigma^2)\right\}$$

- Equivalent to minimize the negative log likelihood (NLL)

$$-\log p(\{y^{(n)}\}_1^N|\{x^{(n)}\}_1^N,\theta,\sigma^2)$$
$$= (N/2)\log\sigma^2 + (2\sigma^2)^{-1}\sum_{n=1}^N (y^{(n)}-\theta^\top x^{(n)})^2 \quad \text{+ const}$$

Maximizing the likelihood

- Likelihood of the training data:

$$p(\{y^{(n)}\}_1^N|\{x^{(n)}\}_1^N,\theta,\sigma^2) = \prod_{n=1}^N p(y^{(n)}|x^{(n)},\theta,\sigma^2)$$
$$= \prod_{n=1}^N \mathcal{N}(y^{(n)}|\theta^\top x^{(n)},\sigma^2)$$
$$= \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-(y^{(n)}-\theta^\top x^{(n)})^2/(2\sigma^2)\right\}$$

- Equivalent to minimize the negative log likelihood (NLL)

$$-\log p(\{y^{(n)}\}_1^N|\{x^{(n)}\}_1^N,\theta,\sigma^2)$$
$$= (N/2)\log\sigma^2 + (2\sigma^2)^{-1}\sum_{n=1}^N (y^{(n)}-\theta^\top x^{(n)})^2 \quad \text{+ const}$$

- The optimal $\hat{\theta}$ can be found without finding $\hat{\sigma}^2$

Maximizing the likelihood

- Likelihood of the training data:

$$p(\{y^{(n)}\}_1^N|\{x^{(n)}\}_1^N,\theta,\sigma^2) = \prod_{n=1}^N p(y^{(n)}|x^{(n)},\theta,\sigma^2)$$
$$= \prod_{n=1}^N \mathcal{N}(y^{(n)}|\theta^\top x^{(n)},\sigma^2)$$
$$= \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-(y^{(n)}-\theta^\top x^{(n)})^2/(2\sigma^2)\right\}$$

- Equivalent to minimize the negative log likelihood (NLL)

$$-\log p(\{y^{(n)}\}_1^N|\{x^{(n)}\}_1^N,\theta,\sigma^2)$$
$$= (N/2)\log\sigma^2 + (2\sigma^2)^{-1}\sum_{n=1}^N (y^{(n)}-\theta^\top x^{(n)})^2 \quad \text{+ const}$$

- The optimal $\hat{\theta}$ can be found without finding $\hat{\sigma}^2$
- $\hat{\theta}$ minimizes $\sum_{n=1}^N (y^{(n)}-\theta^\top x^{(n)})^2$

Maximizing the likelihood

- Likelihood of the training data:

$$p(\{y^{(n)}\}_1^N|\{x^{(n)}\}_1^N,\theta,\sigma^2) = \prod_{n=1}^N p(y^{(n)}|x^{(n)},\theta,\sigma^2)$$
$$= \prod_{n=1}^N \mathcal{N}(y^{(n)}|\theta^\top x^{(n)},\sigma^2)$$
$$= \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-(y^{(n)}-\theta^\top x^{(n)})^2/(2\sigma^2)\right\}$$

- Equivalent to minimize the negative log likelihood (NLL)

$$-\log p(\{y^{(n)}\}_1^N|\{x^{(n)}\}_1^N,\theta,\sigma^2)$$
$$= (N/2)\log\sigma^2 + (2\sigma^2)^{-1}\sum_{n=1}^N (y^{(n)}-\theta^\top x^{(n)})^2 \quad \text{+ const}$$

- The optimal $\hat{\theta}$ can be found without finding $\hat{\sigma}^2$
- $\hat{\theta}$ minimizes $\sum_{n=1}^N (\underbrace{y^{(n)}-\theta^\top x^{(n)}}_{\text{residual (signed)}})^2$

Maximizing the likelihood

- Likelihood of the training data:

$$p(\{y^{(n)}\}_1^N|\{x^{(n)}\}_1^N,\theta,\sigma^2) = \prod_{n=1}^N p(y^{(n)}|x^{(n)},\theta,\sigma^2)$$
$$= \prod_{n=1}^N \mathcal{N}(y^{(n)}|\theta^\top x^{(n)},\sigma^2)$$
$$= \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-(y^{(n)}-\theta^\top x^{(n)})^2/(2\sigma^2)\right\}$$

- Equivalent to minimize the negative log likelihood (NLL)

$$-\log p(\{y^{(n)}\}_1^N|\{x^{(n)}\}_1^N,\theta,\sigma^2)$$
$$= (N/2)\log\sigma^2 + (2\sigma^2)^{-1}\sum_{n=1}^N (y^{(n)}-\theta^\top x^{(n)})^2 \quad \text{+ const}$$

- The optimal $\hat{\theta}$ can be found without finding $\hat{\sigma}^2$
- $\hat{\theta}$ minimizes $\sum_{n=1}^N (\underbrace{y^{(n)}-\theta^\top x^{(n)}}_{\text{residual (signed)}})^2$

- The **residual** is a more general concept: $y^{(n)} - h(x^{(n)})$

Maximizing the likelihood

- Likelihood of the training data:

$$p(\{y^{(n)}\}_1^N|\{x^{(n)}\}_1^N,\theta,\sigma^2) = \prod_{n=1}^N p(y^{(n)}|x^{(n)},\theta,\sigma^2)$$
$$= \prod_{n=1}^N \mathcal{N}(y^{(n)}|\theta^\top x^{(n)},\sigma^2)$$
$$= \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-(y^{(n)}-\theta^\top x^{(n)})^2/(2\sigma^2)\right\}$$

- Equivalent to minimize the negative log likelihood (NLL)

$$-\log p(\{y^{(n)}\}_1^N|\{x^{(n)}\}_1^N,\theta,\sigma^2)$$
$$= (N/2)\log\sigma^2 + (2\sigma^2)^{-1}\sum_{n=1}^N (y^{(n)}-\theta^\top x^{(n)})^2 \quad \text{+ const}$$

- The optimal $\hat{\theta}$ can be found without finding $\hat{\sigma}^2$
- $\hat{\theta}$ minimizes $\sum_{n=1}^N (\underbrace{y^{(n)}-\theta^\top x^{(n)}}_{\text{residual (signed)}})^2$

$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{Residual sum of squares (RSS)}}$

- The **residual** is a more general concept: $y^{(n)} - h(x^{(n)})$

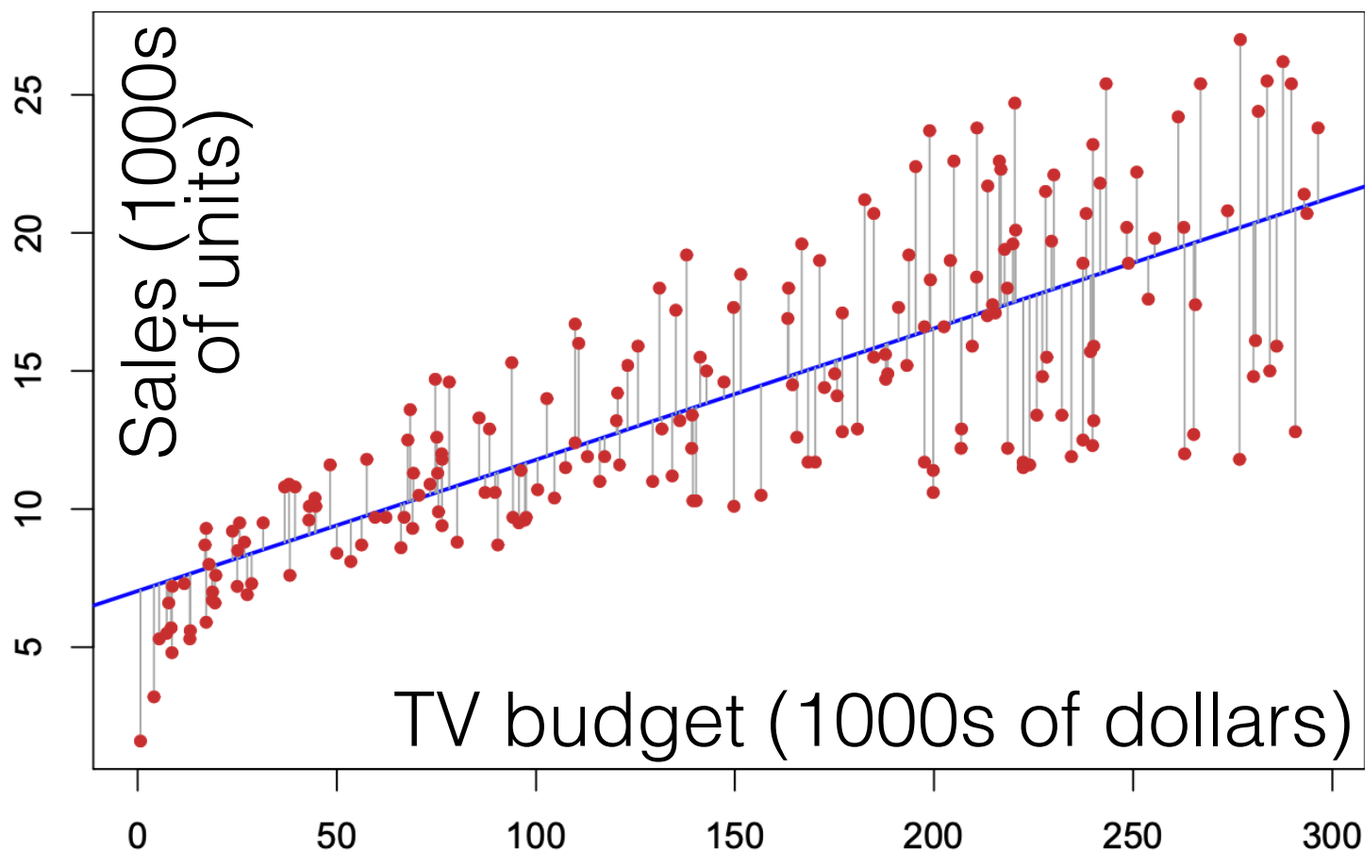
Some visualizations

6a

Some visualizations

- Linear predictor and residuals for one feature (with an intercept)

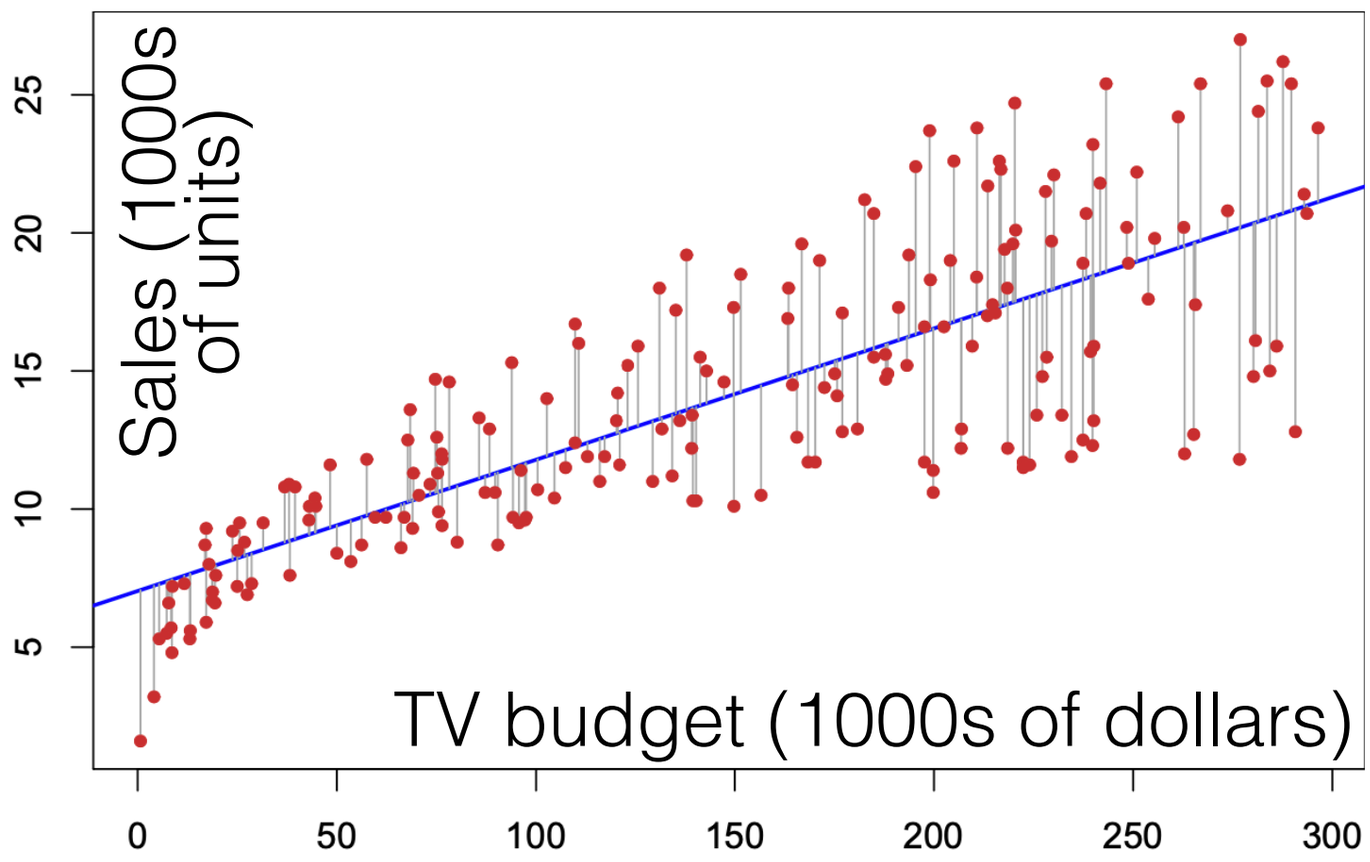
Some visualizations



- Linear predictor and residuals for one feature (with an intercept)

[An Intro to Statistical Learning 2023, Fig 3.1]

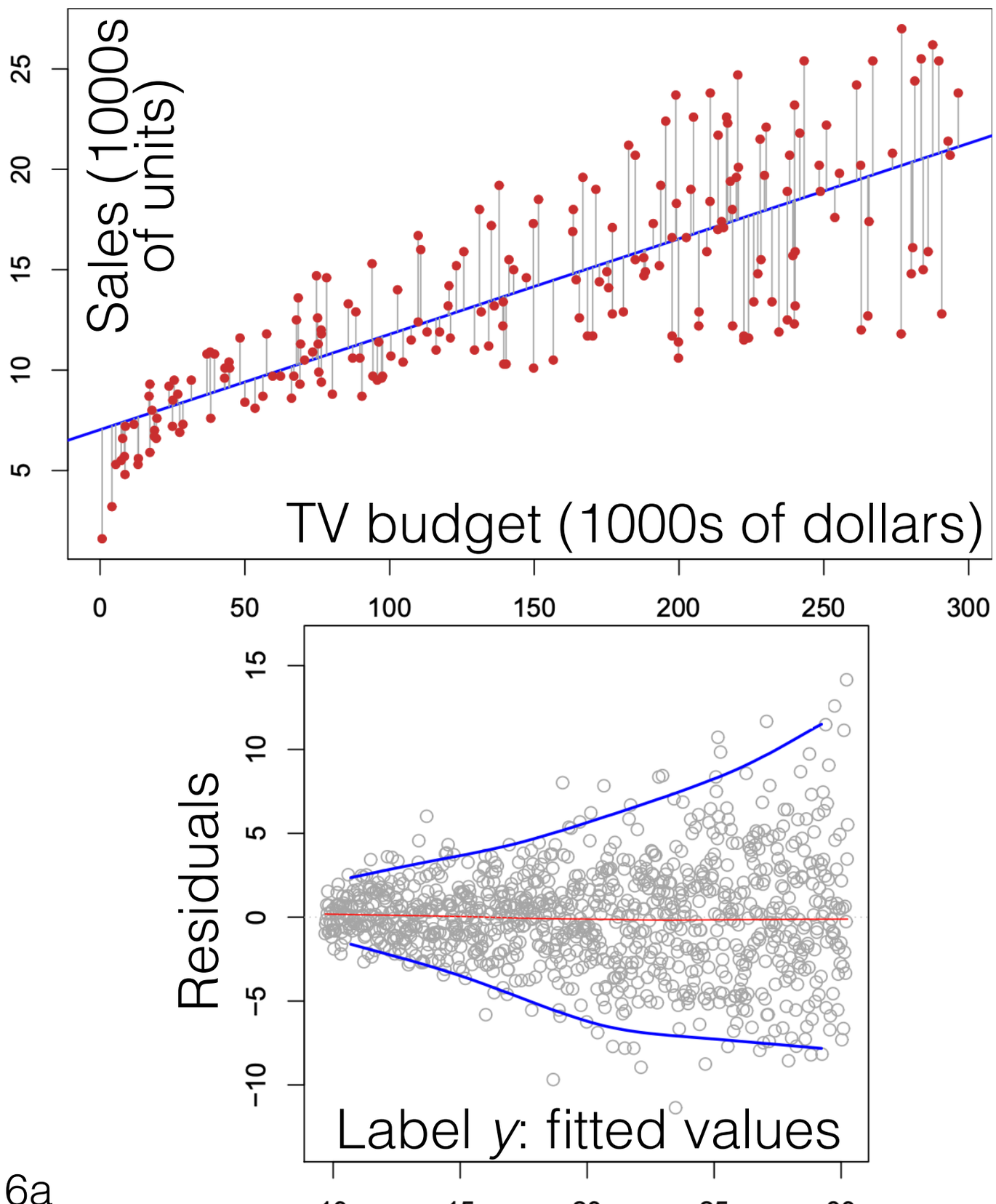
Some visualizations



- Linear predictor and residuals for one feature (with an intercept)
- Can use residual plots of the fitted predictor to check model assumptions

[An Intro to Statistical Learning 2023, Fig 3.1]

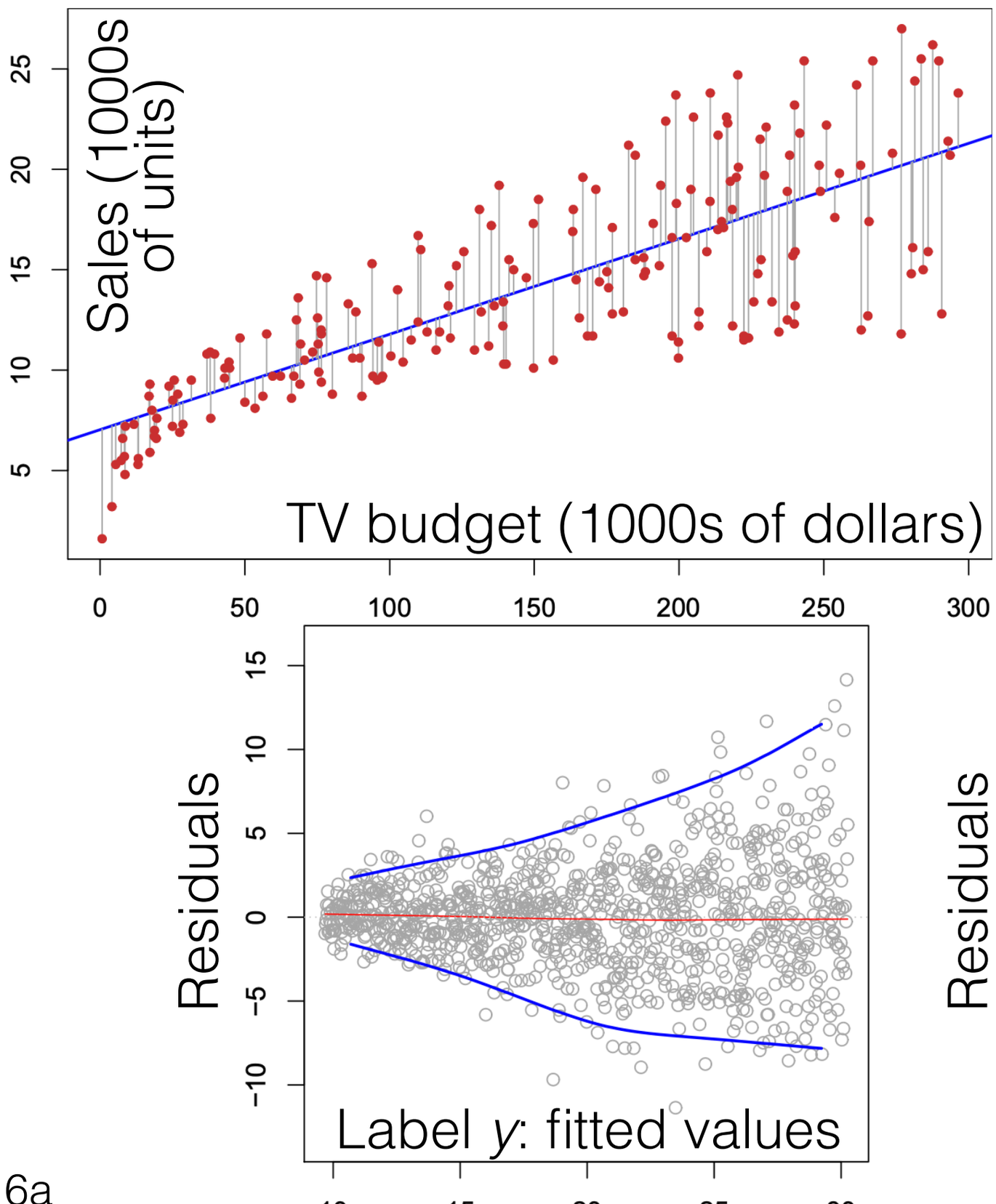
Some visualizations



- Linear predictor and residuals for one feature (with an intercept)
- Can use residual plots of the fitted predictor to check model assumptions

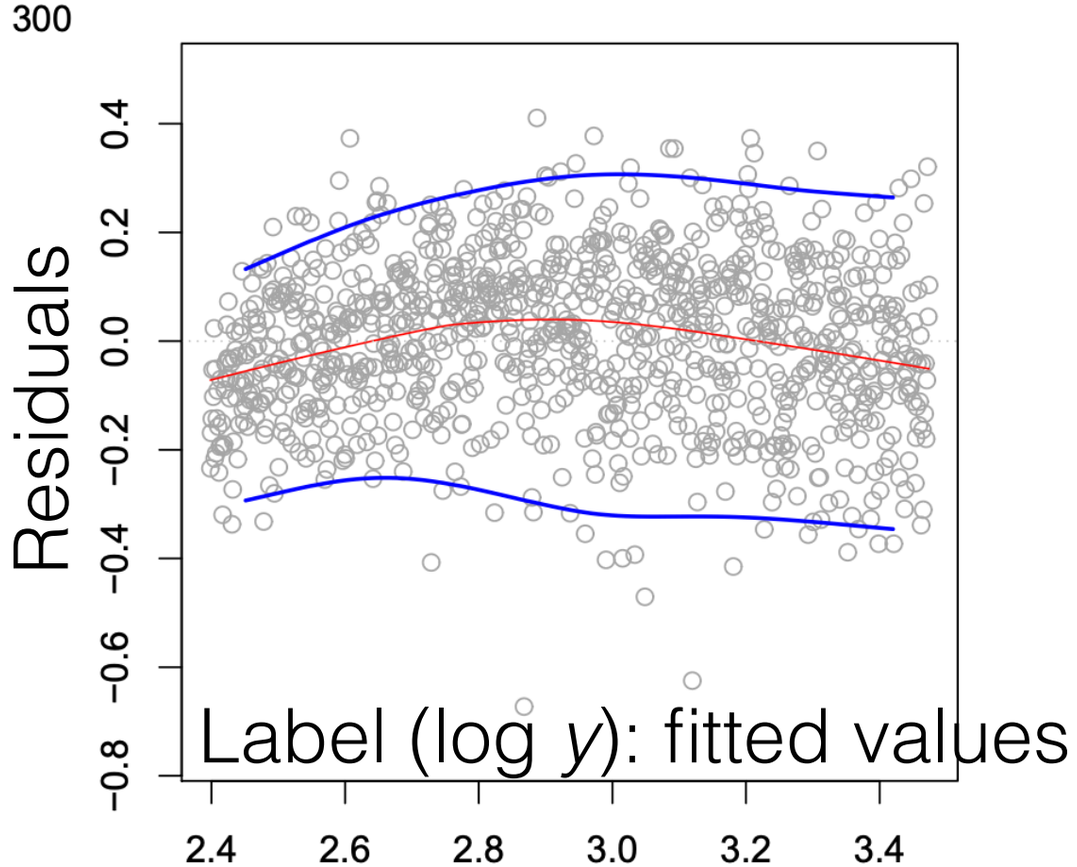
[An Intro to Statistical Learning 2023, Figs 3.1 and 3.11]

Some visualizations

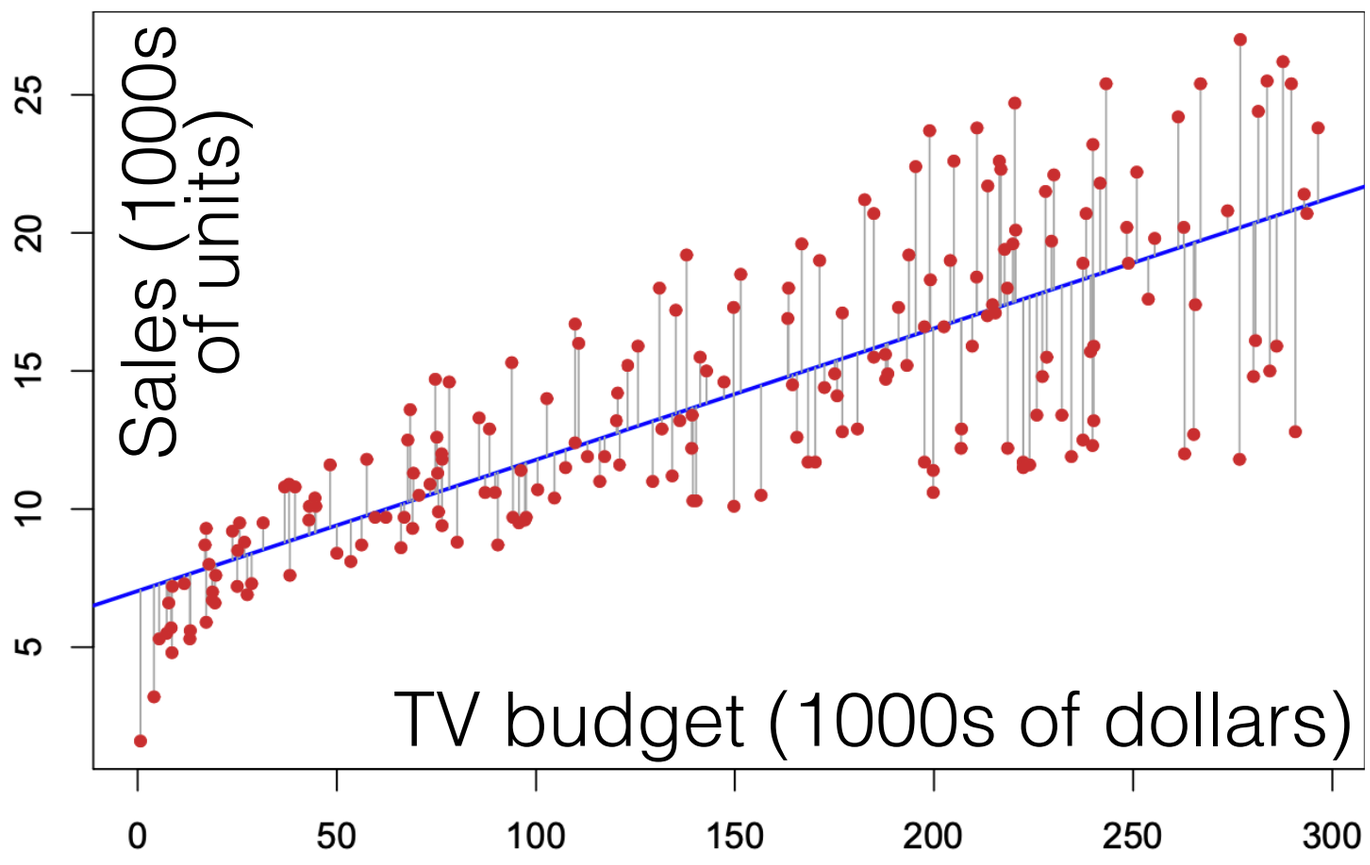


- Linear predictor and residuals for one feature (with an intercept)
- Can use residual plots of the fitted predictor to check model assumptions

[An Intro to Statistical Learning 2023, Figs 3.1 and 3.11]



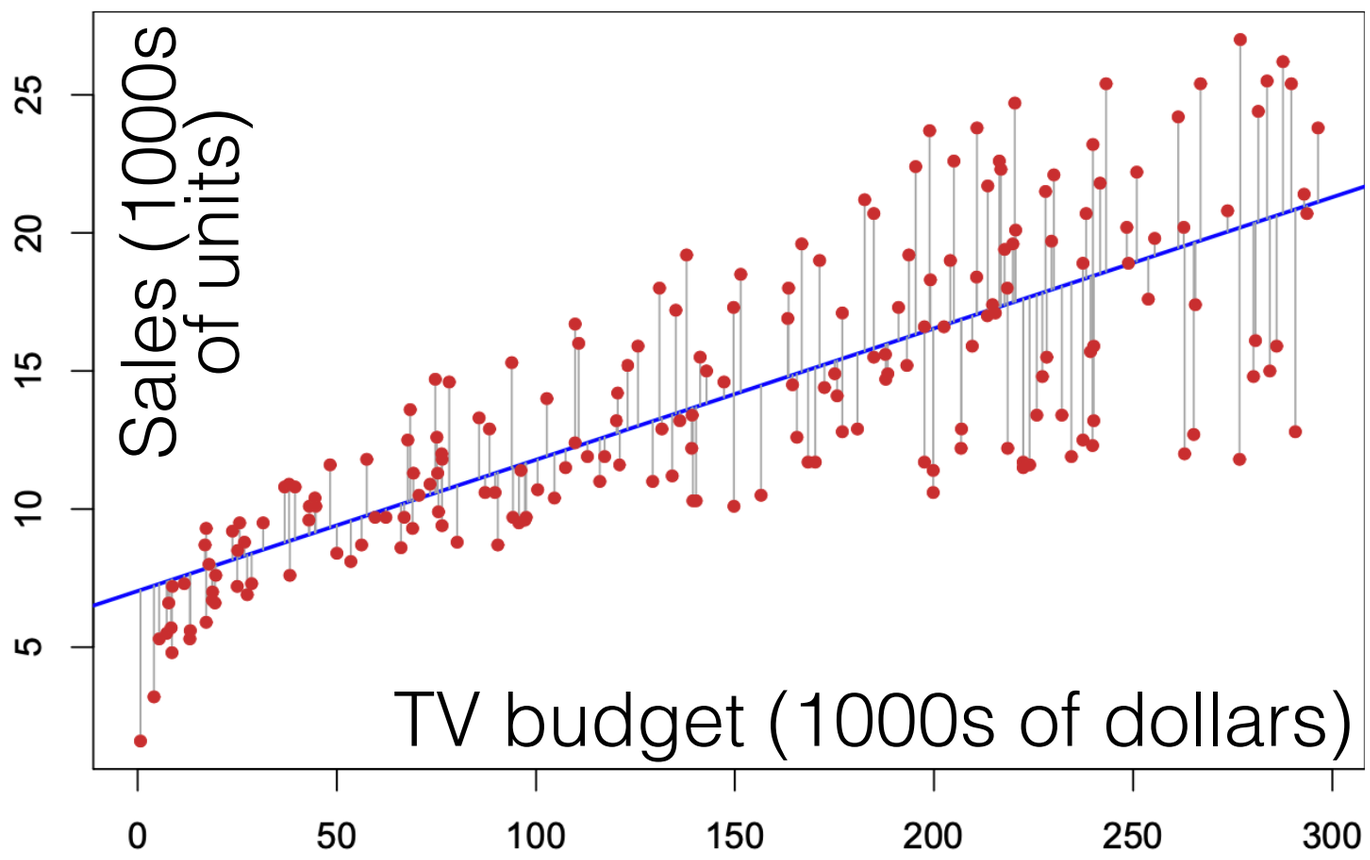
Some visualizations



- Linear predictor and residuals for one feature (with an intercept)
- Can use residual plots of the fitted predictor to check model assumptions

[An Intro to Statistical Learning 2023, Fig 3.1]

Some visualizations

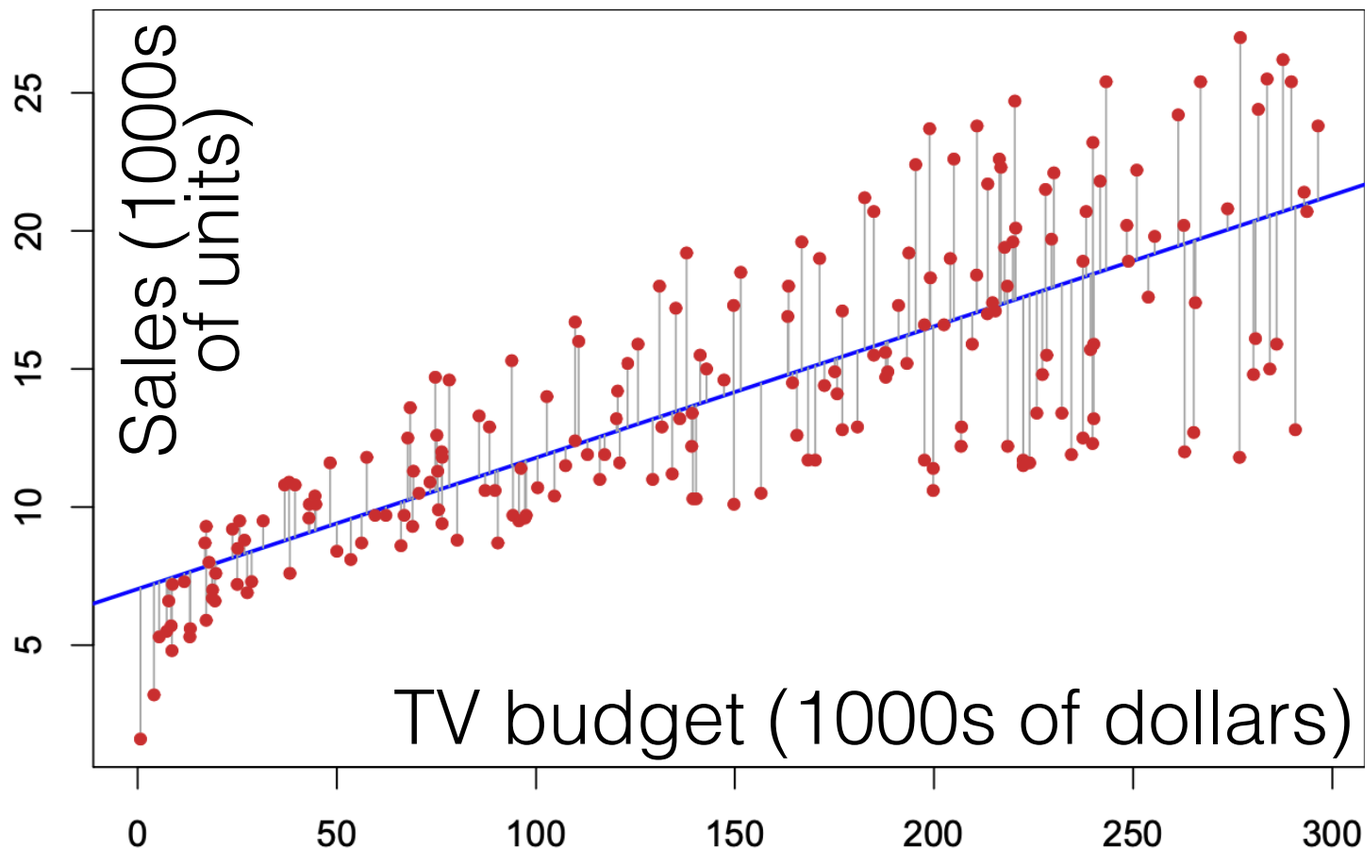


- Linear predictor and residuals for one feature (with an intercept)
- Can use residual plots of the fitted predictor to check model assumptions

[An Intro to Statistical Learning 2023, Fig 3.1]

- Linear predictor & residuals for two features (with an intercept term)

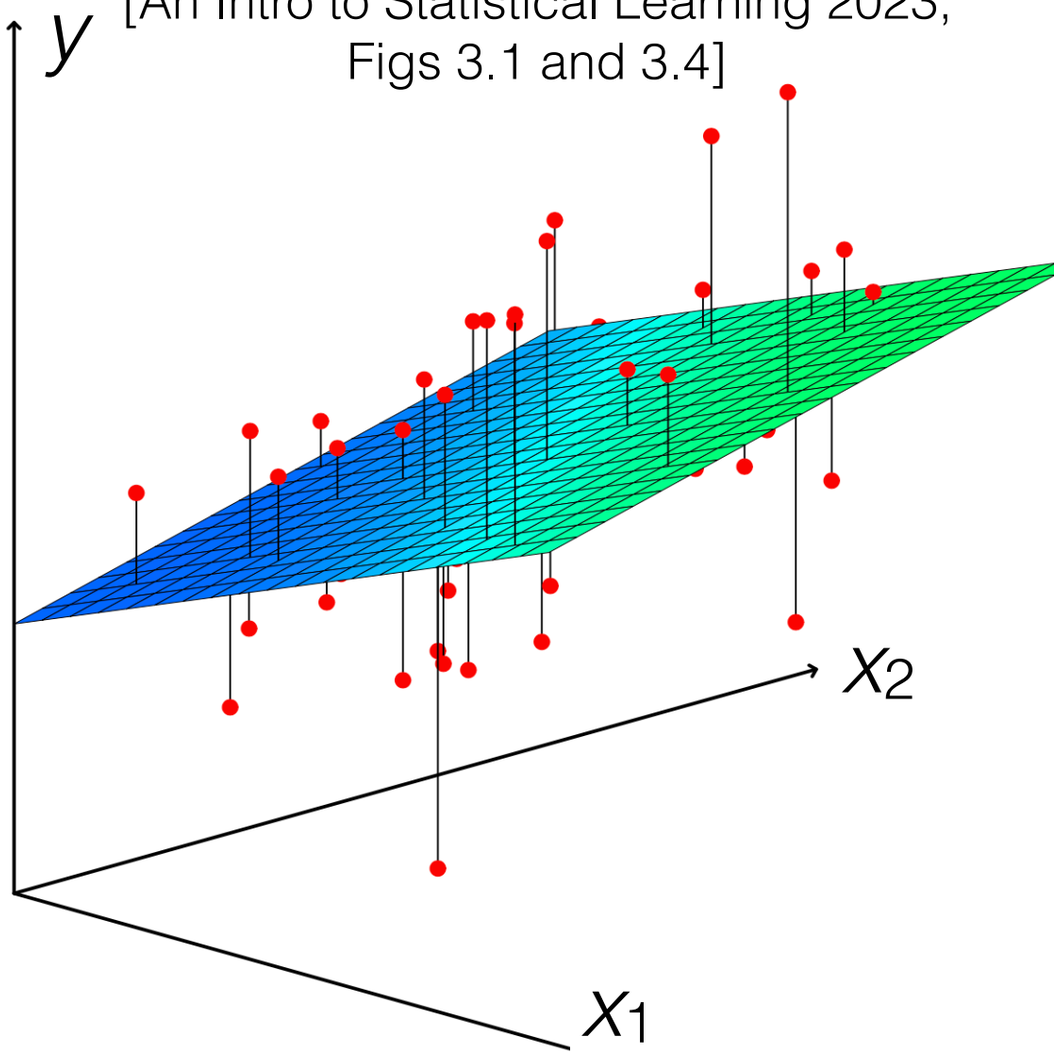
Some visualizations



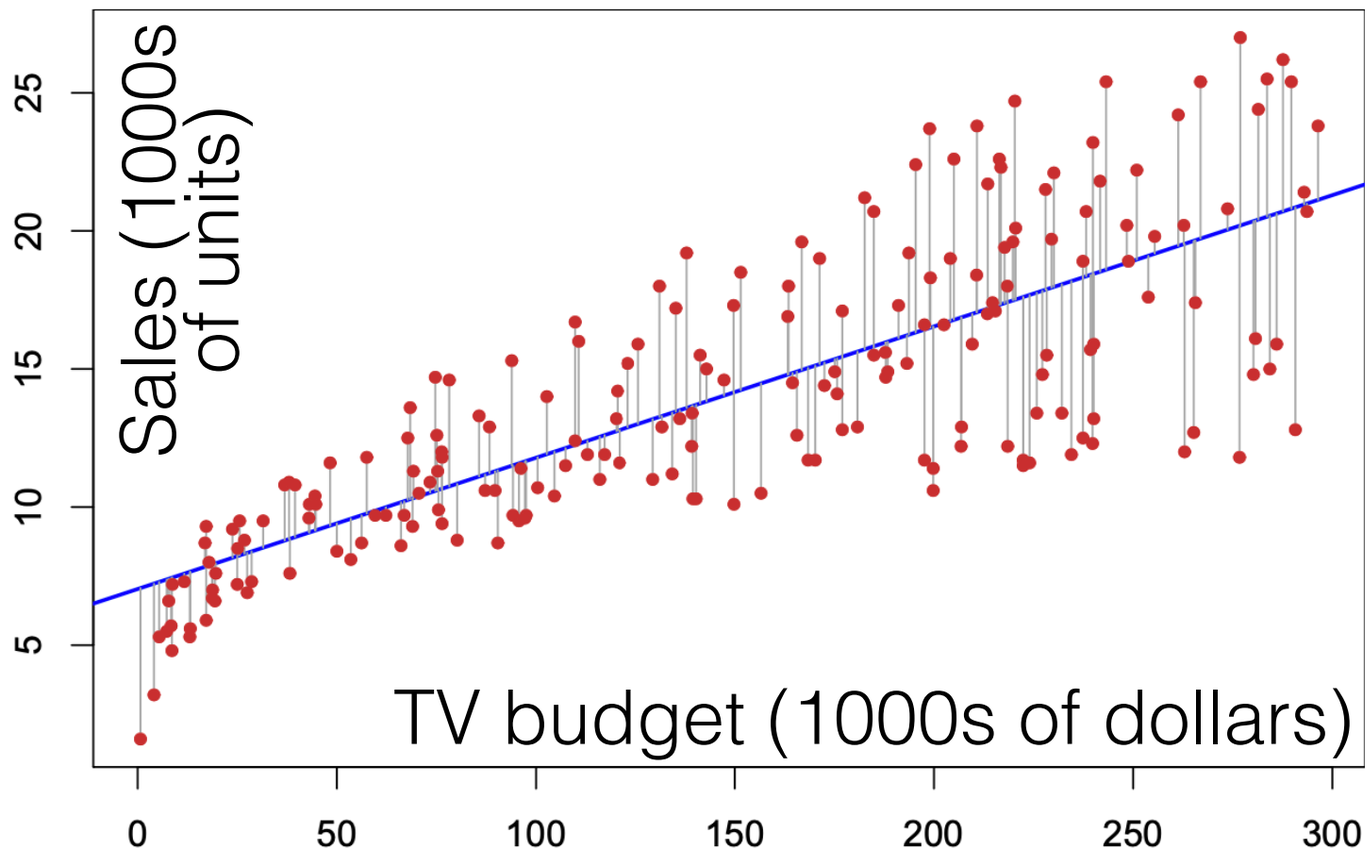
- Linear predictor & residuals for two features (with an intercept term)

- Linear predictor and residuals for one feature (with an intercept)
- Can use residual plots of the fitted predictor to check model assumptions

[An Intro to Statistical Learning 2023, Figs 3.1 and 3.4]



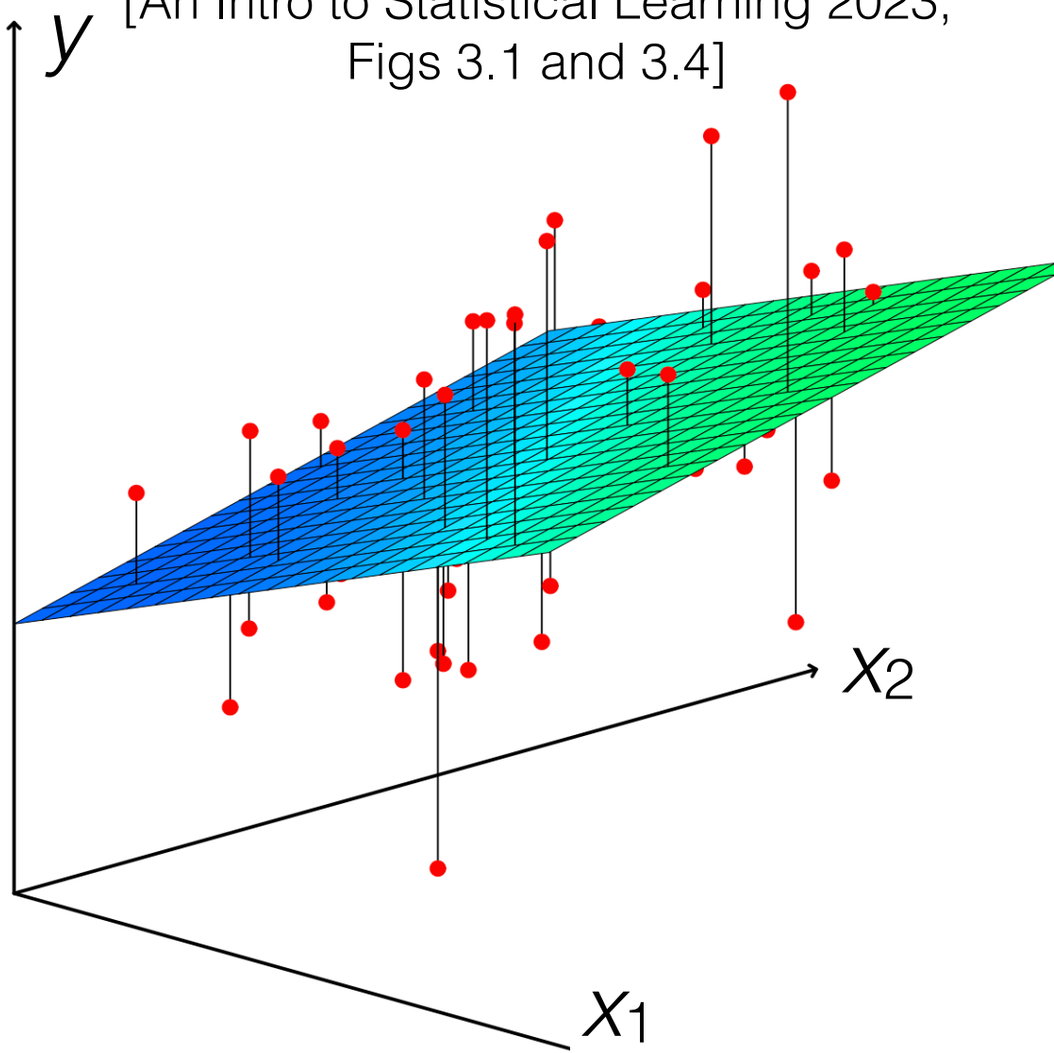
Some visualizations



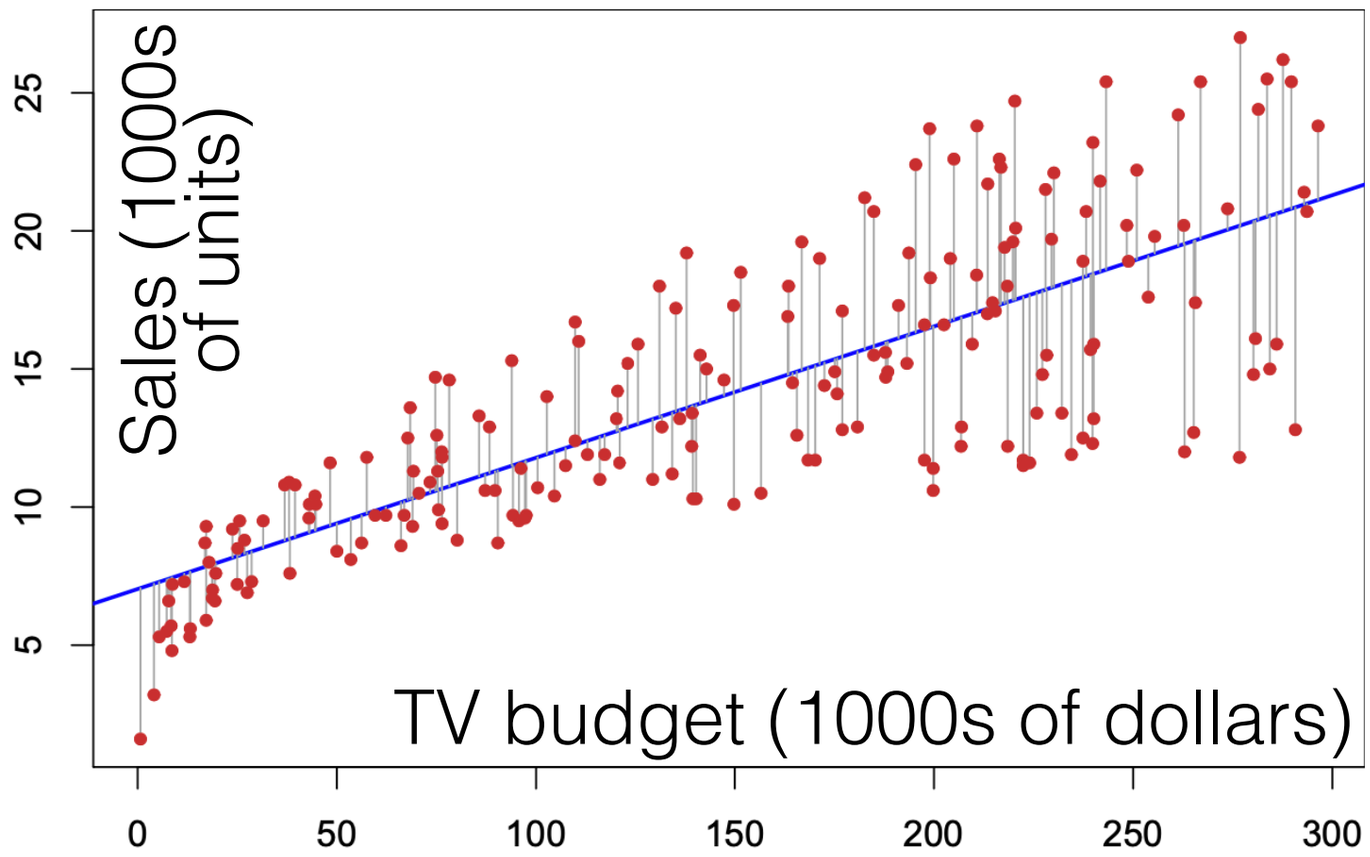
- Linear predictor & residuals for two features (with an intercept term)
- In general, $\theta^\top x$ is a hyperplane (one dim less than (*x,y*) space)

- Linear predictor and residuals for one feature (with an intercept)
- Can use residual plots of the fitted predictor to check model assumptions

[An Intro to Statistical Learning 2023, Figs 3.1 and 3.4]



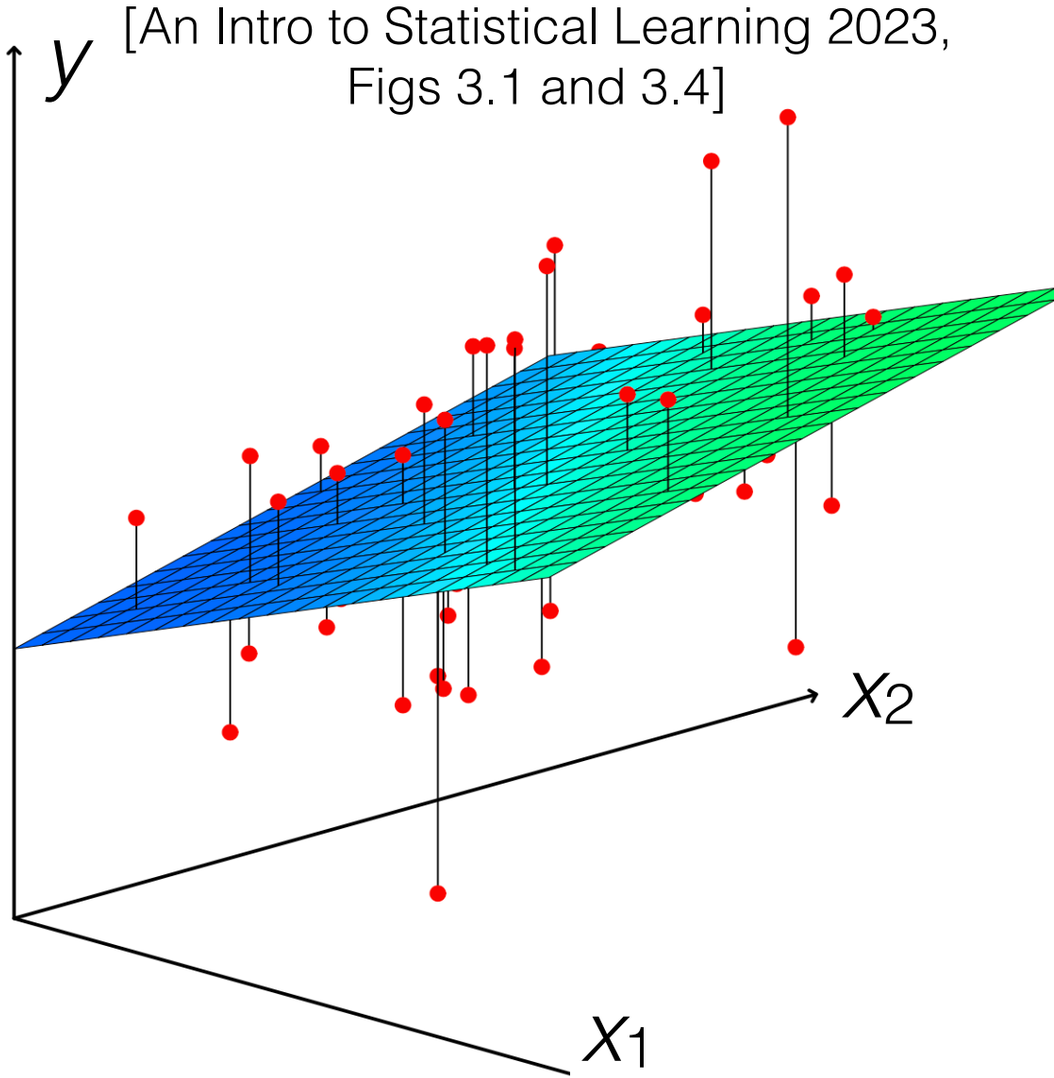
Some visualizations



- Linear predictor & residuals for two features (with an intercept term)
- In general, $\theta^\top x$ is a hyperplane (one dim less than (*x,y*) space)
- Check: regressing label on the features is not equivalent to regressing one feature on label and remaining features

- Linear predictor and residuals for one feature (with an intercept)
- Can use residual plots of the fitted predictor to check model assumptions

[An Intro to Statistical Learning 2023, Figs 3.1 and 3.4]



Empirical risk minimization

Empirical risk minimization

- Empirical risk for square loss:

Empirical risk minimization

- Empirical risk for square loss:

$$\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) =$$

Empirical risk minimization

- Empirical risk for square loss:

$$\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N} (y^{(n)} - h(x^{(n)}))^2$$

Empirical risk minimization

- Empirical risk for square loss:

$$\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N} (y^{(n)} - h(x^{(n)}))^2$$

  - Still not useful if we allow all decision rules (functions) *h*

Empirical risk minimization

- Empirical risk for square loss:

  $\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N} (y^{(n)} - h(x^{(n)}))^2$

  - Still not useful if we allow all decision rules (functions) *h*

- What if we instead allow only linear predictors: $h(x) = \theta^\top x$

Empirical risk minimization

- Empirical risk for square loss:
  $\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - h(x^{(n)}))^2$
  - Still not useful if we allow all decision rules (functions) *h*
- What if we instead allow only linear predictors: $h(x) = \theta^\top x$
  - Then minimizing the empirical risk minimizes
    $\frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - \theta^\top x^{(n)})^2$

Empirical risk minimization

- Empirical risk for square loss:

  $\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - h(x^{(n)}))^2$

  - Still not useful if we allow all decision rules (functions) *h*

- What if we instead allow only linear predictors: $h(x) = \theta^\top x$

  - Then minimizing the empirical risk minimizes

    $\frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - \theta^\top x^{(n)})^2$

    same objective from maximum likelihood! (up to a constant factor)

Empirical risk minimization

- Empirical risk for square loss:

  $\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - h(x^{(n)}))^2$

  - Still not useful if we allow all decision rules (functions) *h*

- What if we instead allow only linear predictors: $h(x) = \theta^\top x$

  - Then minimizing the empirical risk minimizes

    $\frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - \theta^\top x^{(n)})^2$

    same objective from maximum likelihood! (up to a constant factor)

    - relation is specific to this model

Empirical risk minimization

- Empirical risk for square loss:

$$\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N} (y^{(n)} - h(x^{(n)}))^2$$

  - Still not useful if we allow all decision rules (functions) *h*

- What if we instead allow only linear predictors: $h(x) = \theta^\top x$
  - Then minimizing the empirical risk minimizes

$$\underbrace{\frac{1}{N}\sum_{n=1}^{N} (y^{(n)} - \theta^\top x^{(n)})^2}$$

Mean squared error (MSE)

- same objective from maximum likelihood! (up to a constant factor)
- relation is specific to this model

Empirical risk minimization

- Empirical risk for square loss:

  $\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - h(x^{(n)}))^2$

  - Still not useful if we allow all decision rules (functions) *h*

- What if we instead allow only linear predictors: $h(x) = \theta^\top x$

  - Then minimizing the empirical risk minimizes

    $\frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - \theta^\top x^{(n)})^2$

    Mean squared error (MSE)

  - same objective from maximum likelihood! (up to a constant factor)
  - relation is specific to this model

- Notation:

Empirical risk minimization

- Empirical risk for square loss:

$$\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N} (y^{(n)} - h(x^{(n)}))^2$$

  - Still not useful if we allow all decision rules (functions) *h*

- What if we instead allow only linear predictors: $h(x) = \theta^\top x$
  - Then minimizing the empirical risk minimizes

$$\underbrace{\frac{1}{N}\sum_{n=1}^{N} (y^{(n)} - \theta^\top x^{(n)})^2}$$

Mean squared error (MSE)

same objective from maximum likelihood! (up to a constant factor)

- relation is specific to this model

- Notation:

$$X = \begin{pmatrix} (x^{(1)})^\top \\ (x^{(2)})^\top \\ \vdots \\ (x^{(N)})^\top \end{pmatrix}$$

Empirical risk minimization

- Empirical risk for square loss:

$$\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - h(x^{(n)}))^2$$

  - Still not useful if we allow all decision rules (functions) *h*

- What if we instead allow only linear predictors: $h(x) = \theta^\top x$
  - Then minimizing the empirical risk minimizes

$$\underbrace{\frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - \theta^\top x^{(n)})^2}$$

Mean squared error (MSE)

same objective from maximum likelihood! (up to a constant factor)

- relation is specific to this model

- Notation:

$$X = \begin{pmatrix} (x^{(1)})^\top \\ (x^{(2)})^\top \\ \text{Dx1} \\ \vdots \\ (x^{(N)})^\top \end{pmatrix}$$

Empirical risk minimization

- Empirical risk for square loss:

$$\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N} (y^{(n)} - h(x^{(n)}))^2$$

  - Still not useful if we allow all decision rules (functions) *h*

- What if we instead allow only linear predictors: $h(x) = \theta^\top x$
  - Then minimizing the empirical risk minimizes

$$\underbrace{\frac{1}{N}\sum_{n=1}^{N} (y^{(n)} - \theta^\top x^{(n)})^2}$$

Mean squared error (MSE)

same objective from maximum likelihood! (up to a constant factor)

- relation is specific to this model

- Notation:

$$X = \underset{\text{NxD}}{\begin{pmatrix} (x^{(1)})^\top \\ (x^{(2)})^\top \\ \vdots \\ (x^{(N)})^\top \end{pmatrix}}$$

Dx1

Empirical risk minimization

- Empirical risk for square loss:

$$\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N} (y^{(n)} - h(x^{(n)}))^2$$

  - Still not useful if we allow all decision rules (functions) *h*

- What if we instead allow only linear predictors: $h(x) = \theta^\top x$
  - Then minimizing the empirical risk minimizes

$$\underbrace{\frac{1}{N}\sum_{n=1}^{N} (y^{(n)} - \theta^\top x^{(n)})^2}$$

same objective from maximum likelihood! (up to a constant factor)

Mean squared error (MSE)

- relation is specific to this model

- Notation:

$$Y = [y^{(1)}, y^{(2)}, \cdots, y^{(N)}]^\top$$

$$X = \underset{\text{NxD}}{\begin{pmatrix} (x^{(1)})^\top \\ (x^{(2)})^\top \\ \vdots \\ (x^{(N)})^\top \end{pmatrix}}$$

Dx1

Empirical risk minimization

- Empirical risk for square loss:

$$\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - h(x^{(n)}))^2$$

  - Still not useful if we allow all decision rules (functions) *h*

- What if we instead allow only linear predictors: $h(x) = \theta^\top x$
  - Then minimizing the empirical risk minimizes

$$\underbrace{\frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - \theta^\top x^{(n)})^2}$$

Mean squared error (MSE)

same objective from maximum likelihood! (up to a constant factor)

- relation is specific to this model

- Notation:

$$Y = [y^{(1)}, y^{(2)}, \cdots, y^{(N)}]^\top$$

Nx1

$$X = \underset{\text{NxD}}{\begin{pmatrix} (x^{(1)})^\top \\ (x^{(2)})^\top \\ \vdots \\ (x^{(N)})^\top \end{pmatrix}}$$

Dx1

Empirical risk minimization

- Empirical risk for square loss:

$$\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - h(x^{(n)}))^2$$

  - Still not useful if we allow all decision rules (functions) *h*

- What if we instead allow only linear predictors: $h(x) = \theta^\top x$
  - Then minimizing the empirical risk minimizes

$$\underbrace{\frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - \theta^\top x^{(n)})^2}$$

Mean squared error (MSE)

same objective from maximum likelihood! (up to a constant factor)

- relation is specific to this model

- Notation:

$$Y = [y^{(1)}, y^{(2)}, \cdots, y^{(N)}]^\top$$

Nx1

$$X = \underset{\text{NxD}}{\begin{pmatrix} (x^{(1)})^\top \\ (x^{(2)})^\top \\ \vdots \\ (x^{(N)})^\top \end{pmatrix}}$$

Dx1

- Then $\mathrm{RSS}(\theta) =$

Empirical risk minimization

- Empirical risk for square loss:

$$\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - h(x^{(n)}))^2$$

  - Still not useful if we allow all decision rules (functions) *h*

- What if we instead allow only linear predictors: $h(x) = \theta^\top x$
  - Then minimizing the empirical risk minimizes

$$\underbrace{\frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - \theta^\top x^{(n)})^2}$$

Mean squared error (MSE)

same objective from maximum likelihood! (up to a constant factor)

- relation is specific to this model

- Notation:

$$Y = [y^{(1)}, y^{(2)}, \cdots, y^{(N)}]^\top$$

Nx1

$$X = \underset{\text{NxD}}{\begin{pmatrix} (x^{(1)})^\top \\ (x^{(2)})^\top \\ \vdots \\ (x^{(N)})^\top \end{pmatrix}}$$

Dx1

- Then $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$

Empirical risk minimization

- Empirical risk for square loss:

$$\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - h(x^{(n)}))^2$$

  - Still not useful if we allow all decision rules (functions) *h*

- What if we instead allow only linear predictors: $h(x) = \theta^\top x$
  - Then minimizing the empirical risk minimizes

$$\underbrace{\frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - \theta^\top x^{(n)})^2}$$

Mean squared error (MSE)

same objective from maximum likelihood! (up to a constant factor)

- relation is specific to this model

- Notation:

$$X = \underset{\text{NxD}}{\begin{pmatrix} \underset{\text{Dx1}}{(x^{(1)})^\top} \\ (x^{(2)})^\top \\ \vdots \\ (x^{(N)})^\top \end{pmatrix}} \qquad Y = \underset{\text{Nx1}}{[y^{(1)}, y^{(2)}, \cdots, y^{(N)}]^\top}$$

  - Then $\underset{\text{NxD}}{\mathrm{RSS}(\theta)} = \underset{\text{Nx1}}{(X\theta - Y)^\top (X\theta - Y)}$

Empirical risk minimization

- Empirical risk for square loss:

$$\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - h(x^{(n)}))^2$$

  - Still not useful if we allow all decision rules (functions) *h*

- What if we instead allow only linear predictors: $h(x) = \theta^\top x$
  - Then minimizing the empirical risk minimizes

$$\underbrace{\frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - \theta^\top x^{(n)})^2}$$

Mean squared error (MSE)

same objective from maximum likelihood! (up to a constant factor)

- relation is specific to this model

- Notation:

$$Y = [y^{(1)}, y^{(2)}, \cdots, y^{(N)}]^\top$$

Nx1

$$X = \underset{\text{NxD}}{\begin{pmatrix} (x^{(1)})^\top \\ (x^{(2)})^\top \\ \vdots \\ (x^{(N)})^\top \end{pmatrix}}$$

Dx1

- Then $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$

NxD Dx1 Nx1

Empirical risk minimization

- Empirical risk for square loss:

$$\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - h(x^{(n)}))^2$$

  - Still not useful if we allow all decision rules (functions) *h*

- What if we instead allow only linear predictors: $h(x) = \theta^\top x$
  - Then minimizing the empirical risk minimizes

$$\underbrace{\frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - \theta^\top x^{(n)})^2}$$

same objective from maximum likelihood! (up to a constant factor)

Mean squared error (MSE)

- relation is specific to this model

- Notation:

$$Y = [y^{(1)}, y^{(2)}, \cdots, y^{(N)}]^\top$$

Nx1

$$X = \underset{\text{NxD}}{\begin{pmatrix} (x^{(1)})^\top \\ (x^{(2)})^\top \\ \vdots \\ (x^{(N)})^\top \end{pmatrix}}$$

Dx1

- Then $\underset{\text{Nx1}}{\text{RSS}(\theta)} = \underbrace{(X\theta - Y)^\top}_{\text{NxD Dx1}}\underbrace{(X\theta - Y)}_{\text{Nx1}}$

Empirical risk minimization

- Empirical risk for square loss:

$$\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - h(x^{(n)}))^2$$

  - Still not useful if we allow all decision rules (functions) *h*

- What if we instead allow only linear predictors: $h(x) = \theta^\top x$
  - Then minimizing the empirical risk minimizes

$$\underbrace{\frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - \theta^\top x^{(n)})^2}$$

Mean squared error (MSE)

same objective from maximum likelihood! (up to a constant factor)

- relation is specific to this model

- Notation:

$$X = \underset{\text{NxD}}{\begin{pmatrix} (x^{(1)})^\top \\ (x^{(2)})^\top \\ \vdots \\ (x^{(N)})^\top \end{pmatrix}}$$

Dx1

$$Y = \underset{\text{Nx1}}{[y^{(1)}, y^{(2)}, \cdots, y^{(N)}]^\top}$$

- Then $\underset{\text{NxD}\ \ \text{Dx1}\ \ \text{Nx1}}{\text{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)}$

$\underbrace{\qquad}_{\text{1xN}}$ $\underbrace{\qquad}_{\text{Nx1}}$

Empirical risk minimization

- Empirical risk for square loss:

$$\frac{1}{N}\sum_{n=1}^{N} L(y^{(n)}, h(x^{(n)})) = \frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - h(x^{(n)}))^2$$

  - Still not useful if we allow all decision rules (functions) *h*

- What if we instead allow only linear predictors: $h(x) = \theta^\top x$
  - Then minimizing the empirical risk minimizes

$$\underbrace{\frac{1}{N}\sum_{n=1}^{N}(y^{(n)} - \theta^\top x^{(n)})^2}$$

same objective from maximum likelihood! (up to a constant factor)

Mean squared error (MSE)

- relation is specific to this model

- Notation:

$$Y = [y^{(1)}, y^{(2)}, \cdots, y^{(N)}]^\top$$

Nx1

$$X = \underset{\text{NxD}}{\begin{pmatrix} (x^{(1)})^\top \\ (x^{(2)})^\top \\ \vdots \\ (x^{(N)})^\top \end{pmatrix}}$$

Dx1

- Then $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$

NxD Dx1 Nx1

1xN Nx1

1x1

Let’s optimize!

Let’s optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions:

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) =$

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2X^\top (X\theta - Y)$

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2X^\top (X\theta - Y)$

check you can derive!

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2X^\top (X\theta - Y) \overset{\text{set}}{=} 0$

check you can derive!

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with dth element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (d,d') element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2X^\top (X\theta - Y) \overset{\text{set}}{=} 0$

check you can derive!

Dx1

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2\underbrace{X^\top}_{\text{DxN}}(X\theta - Y) \overset{\text{set}}{=} 0$

check you can derive!

Dx1

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2\underbrace{X^\top}_{\text{DxN}}\underbrace{(X\theta - Y)}_{\text{Nx1}} \overset{\text{set}}{=} \underbrace{0}_{\text{Dx1}}$

check you can derive!

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2\underbrace{X^\top}_{\text{DxN}}\underbrace{(X\theta - Y)}_{\text{Nx1}} \overset{\text{set}}{=} \underbrace{0}_{\text{Dx1}}$

check you can derive!

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2\underbrace{X^\top}_{\text{DxN}}\underbrace{(X\theta - Y)}_{\text{Nx1}} \overset{\text{set}}{=} \underbrace{0}_{\text{Dx1}}$

check you can derive!

  - Second-order conditions:

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2\underbrace{X^\top}_{\text{DxN}}\underbrace{(X\theta - Y)}_{\text{Nx1}} \overset{\text{set}}{=} \underbrace{0}_{\text{Dx1}}$

check you can derive!

  - Second-order conditions: $\nabla^2 \mathrm{RSS}(\theta) =$

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2\underbrace{X^\top}_{\text{DxN}}\underbrace{(X\theta - Y)}_{\text{Nx1}} \overset{\text{set}}{=} \underbrace{0}_{\text{Dx1}}$

check you can derive!

  - Second-order conditions: $\nabla^2 \mathrm{RSS}(\theta) = 2X^\top X$

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2\underbrace{X^\top}_{\text{DxN}}\underbrace{(X\theta - Y)}_{\text{Nx1}} \overset{\text{set}}{=} \underbrace{0}_{\text{Dx1}}$

check you can derive!

  - Second-order conditions: $\nabla^2 \mathrm{RSS}(\theta) = 2\underbrace{X^\top X}_{\text{DxD}}$

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2\underbrace{X^\top}_{\text{DxN}}\underbrace{(X\theta - Y)}_{\text{Nx1}} \overset{\text{set}}{=} \underbrace{0}_{\text{Dx1}}$

check you can derive!

  - Second-order conditions: $\nabla^2 \mathrm{RSS}(\theta) = 2\underbrace{X^\top X}_{\text{DxD}}$

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2\underbrace{X^\top}_{\text{DxN}}\underbrace{(X\theta - Y)}_{\text{Nx1}} \overset{\text{set}}{=} \underbrace{0}_{\text{Dx1}}$

check you can derive!

  - Second-order conditions: $\nabla^2 \mathrm{RSS}(\theta) = 2\underbrace{X^\top X}_{\text{DxD}} \overset{\text{set}}{>} 0$ is positive definite

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2\underbrace{X^\top}_{\text{DxN}}\underbrace{(X\theta - Y)}_{\text{Nx1}} \overset{\text{set}}{=} 0$ (Dx1; Dx1)

check you can derive!

  - Second-order conditions: $\nabla^2 \mathrm{RSS}(\theta) = 2\underbrace{X^\top X}_{\text{DxD}} \overset{\text{set}}{>} 0$ (is positive definite)
  - For now, suppose *N*>*D* & *X* is full rank

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2\underbrace{X^\top}_{\text{DxN}}\underbrace{(X\theta - Y)}_{\text{Nx1}} \overset{\text{set}}{=} \underbrace{0}_{\text{Dx1}}$

check you can derive!

  - Second-order conditions: $\nabla^2 \mathrm{RSS}(\theta) = 2\underbrace{X^\top X}_{\text{DxD}} \overset{\text{set}}{>} 0$ is positive definite
- For now, suppose *N*>*D* & *X* is full rank
  - I.e. *X* has rank *D*, i.e. columns linearly independent

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2\underbrace{X^\top}_{\text{DxN}}\underbrace{(X\theta - Y)}_{\text{Nx1}} \overset{\text{set}}{=} \underbrace{0}_{\text{Dx1}}$

check you can derive!

  - Second-order conditions: $\nabla^2 \mathrm{RSS}(\theta) = 2\underbrace{X^\top X}_{\text{DxD}} \overset{\text{set}}{>} 0$ is positive definite
- For now, suppose *N*>*D* & *X* is full rank
  - I.e. *X* has rank *D*, i.e. columns linearly independent
  - Then $X^\top X$ is positive definite and invertible

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2\underbrace{X^\top}_{\text{DxN}}\underbrace{(X\theta - Y)}_{\text{Nx1}} \overset{\text{set}}{=} \underbrace{0}_{\text{Dx1}}$

check you can derive!

  - Second-order conditions: $\nabla^2 \mathrm{RSS}(\theta) = 2\underbrace{X^\top X}_{\text{DxD}} \overset{\text{set}}{>} 0$ is positive definite
- For now, suppose *N*>*D* & *X* is full rank
  - I.e. *X* has rank *D*, i.e. columns linearly independent
  - Then $X^\top X$ is positive definite and invertible
  - And we can solve the first-order conditions:

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2\underbrace{X^\top}_{\text{DxN}}\underbrace{(X\theta - Y)}_{\text{Nx1}} \overset{\text{set}}{=} 0$ (Dx1; Dx1)

check you can derive!

  - Second-order conditions: $\nabla^2 \mathrm{RSS}(\theta) = 2\underbrace{X^\top X}_{\text{DxD}} \overset{\text{set}}{>} 0$ (is positive definite)
- For now, suppose *N*>*D* & *X* is full rank
  - I.e. *X* has rank *D*, i.e. columns linearly independent
  - Then $X^\top X$ is positive definite and invertible
  - And we can solve the first-order conditions:

$$X^\top X \hat{\theta} = X^\top Y$$

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2\underbrace{X^\top}_{\text{DxN}}\underbrace{(X\theta - Y)}_{\text{Nx1}} \overset{\text{set}}{=} 0$ (Dx1; Dx1)

check you can derive!

  - Second-order conditions: $\nabla^2 \mathrm{RSS}(\theta) = 2\underbrace{X^\top X}_{\text{DxD}} \overset{\text{set}}{>} 0$ (is positive definite)
- For now, suppose *N*>*D* & *X* is full rank
  - I.e. *X* has rank *D*, i.e. columns linearly independent
  - Then $X^\top X$ is positive definite and invertible
  - And we can solve the first-order conditions:

$$X^\top X \hat{\theta} = X^\top Y$$
$$\hat{\theta} = (X^\top X)^{-1} X^\top Y$$

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2\underbrace{X^\top}_{\text{DxN}}\underbrace{(X\theta - Y)}_{\text{Nx1}} \overset{\text{set}}{=} \underbrace{0}_{\text{Dx1}}$

check you can derive!

  - Second-order conditions: $\nabla^2 \mathrm{RSS}(\theta) = 2\underbrace{X^\top X}_{\text{DxD}} \overset{\text{set}}{>} 0$ is positive definite
- For now, suppose *N*>*D* & *X* is full rank
  - I.e. *X* has rank *D*, i.e. columns linearly independent
  - Then $X^\top X$ is positive definite and invertible
  - And we can solve the first-order conditions:

$$X^\top X \hat{\theta} = X^\top Y$$
$$\hat{\theta} = (X^\top X)^{-1} X^\top Y$$

"ordinary least squares" (OLS)

Let's optimize!

- We want to minimize $\mathrm{RSS}(\theta) = (X\theta - Y)^\top (X\theta - Y)$
  - Notation: for function $f(\theta)$, the gradient $\nabla_\theta f$ is the Dx1 vector with *d*th element: $\partial f / \partial \theta_d$
  - The Hessian $\nabla^2_\theta f$ is the DxD matrix with (*d*,*d'*) element $\frac{\partial^2 f}{\partial \theta_d \partial \theta_{d'}}$
  - First-order conditions: $\nabla \mathrm{RSS}(\theta) = 2\underbrace{X^\top}_{\text{DxN}}\underbrace{(X\theta - Y)}_{\text{Nx1}} \overset{\text{set}}{=} \underbrace{0}_{\text{Dx1}}$

check you can derive!

  - Second-order conditions: $\nabla^2 \mathrm{RSS}(\theta) = 2\underbrace{X^\top X}_{\text{DxD}} \overset{\text{set}}{>} 0$ is positive definite
- For now, suppose *N*>*D* & *X* is full rank
  - I.e. *X* has rank *D*, i.e. columns linearly independent
  - Then $X^\top X$ is positive definite and invertible
  - And we can solve the first-order conditions:

$$X^\top X \hat{\theta} = X^\top Y$$
$$\hat{\theta} = (X^\top X)^{-1} X^\top Y$$

"ordinary least squares" (OLS)

- A closed-form solution isn't always better
  - Matrix inversion can be expensive for large dimensions

References (1/1)

Castro Torres, Andrés F., and Aliakbar Akbaritabar. "The use of linear models in quantitative research." *Quantitative Science Studies* (2024): 1-21.

Chen, Danqi, Jason Bolton, and Christopher D. Manning. "A thorough examination of the CNN/ Daily Mail reading comprehension task." *ACL* (2016).

Lipton, Zachary C., and Jacob Steinhardt. "Troubling Trends in Machine Learning Scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research." Queue 17.1 (2019): 45-77.

Salganik, Matthew J., et al. "Measuring the predictability of life outcomes with a scientific mass collaboration." Proceedings of the National Academy of Sciences 117.15 (2020): 8398-8403.