EML'24 – Lecture 3

Linear Regression II

ISLR 3, ESL 3

Prof. Isabel Valera

31 October 2024

Saarland Informatics Campus

Four Important Questions

1. Is at least one predictor useful?
2. Which subset of predictors is useful?
3. How well does the model fit the data?
4. How accurately can we predict the response?

Question: Is at Least one Predictor Useful?

To tell whether at least one predictor is useful we have to test

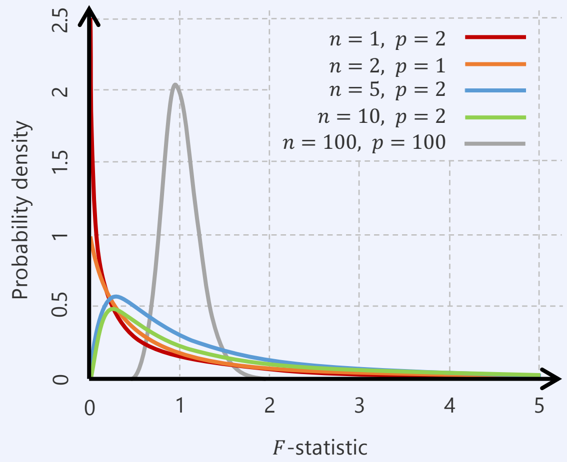
- $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ vs. H_a: at least one β_i is non-zero
- we can test this using the F-statistic

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

- under linear assumptions, we have $\mathrm{E}[RSS/(n-p-1)] = \sigma^2$
- if H_0 is true then $\mathrm{E}[(TSS - RSS)/p] = \sigma^2$
  else $\mathrm{E}[(TSS - RSS)/p] > \sigma^2$

The F-statistic is 1 if H_0 is true, and greater than 1 otherwise

- for the advertising data, the F-statistic is 570
- in general, the F-statistic follows an F-distribution



Question Which Subset of Predictors is Useful?

To test subsets of predictors we can again define a hypothesis test

- i.e. we can test whether features are useful in addition to some set of predictors
- $H_0: \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$, i.e. we test if the last q predictors in the list are (un)informative

The corresponding F-statistic is

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

- RSS_0 is the RSS of a model that includes all except the last q variables

	Coefficient	Std. error	t-statistic	p-value
intercept	2.939	0.3119	9.42	<0.0001
TV	0.046	0.0014	32.81	<0.0001
radio	0.189	0.0086	21.89	<0.0001
newspaper	-0.001	0.0059	-0.18	0.8599

The reported t-statistics in the table are the square-roots of the F-statistic

- they measure the added effect of that variable when all other variables are included in the model
- e.g. **newspaper** adds no effect to a model that includes both **TV** and **radio**

What if we have many predictors to choose from?

In high-dimensional settings we cannot restrict ourselves to p-values of individual variables

- assume $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ with $p = 100$ to be true
- we generate a random response, so, no variable is associated with it
- we are practically guaranteed to find a result with a 'significant' result
- due to multiple testing just by chance 5% of the p-values will be below 5%
- the $\boldsymbol{t}$-statistic does not adjust for number of predictors, but the $\boldsymbol{F}$-statistic does

If $p > n$ this does not help, as we have too few observations to fit all parameters

- *oh noes, what now?* wait till Chapter 6

Preview Selecting Important Variables

Often, the outcome is only dependent on a few variables

- finding those variables is the variable selection or feature selection problem
- Chapter 6 discusses this in detail. Here we give a preview.

Preview Selecting Important Variables

Best subset selection

Try all subsets of variables, here

- {}, {**TV**},{**radio**},{**newspaper**}, {**TV**, **radio**},{**TV**, **newspaper**}, {**radio**, **newspaper**}, {**TV**, **radio**, **newspaper**}
- there are 2^p subsets
- $p = 30, \; then \; 2^{30} = 1{,}073{,}741{,}824$ models

How does one rate the performance of a model?

- not via the training error!
- need methods to assess test error

Preview Selecting Important Variables

Forward Selection

- begin with the null model over no variables
- fit p models, one with each single variable
- select the model with lowest RSS
- try adding all of the remaining $p - 1$ variables into this model
- pick the one with the lowest RSS
- continue, until a stopping criterion is fulfilled

Backward selection

- begin with the full model over all variables
- remove the variable with the largest p-value according to the F-statistic
- continue, until a stopping criterion is fulfilled

Mixed Selection

- begin with the null model over no variables
- add variables to the model until the added variable becomes insignificant
- remove variables until there is no insignificant variable in the model
- continue until all variables in the model are significant, and all variables outside are not

Question: How Well Does the Model Fit the Data?

The most common numerical measures for model fit are RSE and R^2

- for univariate regression, $R^2 = Corr(X, \hat{Y})^2$
- for multivariate regression, $R^2 = Corr(Y, \hat{Y})^2$
- among all linear models the full linear model maximizes correlation

Reminder:
$R^2 = 1 - \frac{RSS}{TSS}$; where

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

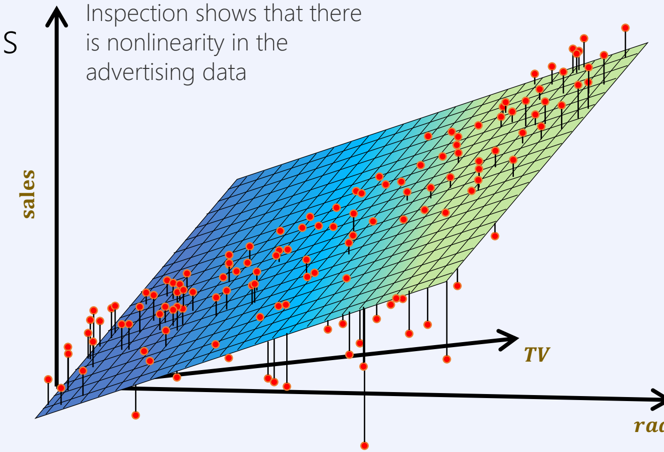
$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$

$(Training) R^2$ monotonically increases when we add variables

- even if these are only weakly associated with the output
- really, we need to consider test error (Chapter 5)

	RSE	R^2	F-statistic
Full Model	1.681	0.8972	570
TV, radio	1.686	0.89719	
TV	3.26	0.612	312.6

Inspection shows that there is nonlinearity in the advertising data



Inspect the data!

Question: How Accurately can we Predict?

1. Predict outcome based on trained linear model
   1. inaccuracy of **coefficient estimates** are related to the reducible error
   2. we compute **confidence intervals** for coefficients and for the output
2. When the relationship between input and output is non-linear, any linear model will incur a bias and the **reducible error** can be further reduced with a non-linear model!

Even when we know the true relationship, we can never remove the irreducible error

- confidence intervals relate to the variability of an estimate **over many samples**
- prediction intervals relate to the variability of an estimate **for a given sample**
- **prediction intervals** are hence always wider than the confidence intervals, as it accounts for the aleatoric error
- for example, on the advertising data
  - $\mathbf{TV} = \$100{,}000$, $\mathbf{radio} = \$20{,}000$
  - 95%-confidence interval $\mathbf{sales} \in [10985, 11528]$
  - 95%-prediction interval $\mathbf{sales} \in [7930, 14580]$

confidence interval $\mathrm{E}[y|x] = \hat{y} \quad \pm t_{\alpha/2,n-2}\sqrt{MSE}\sqrt{\frac{1}{n} + \frac{(x-x)^2}{\sum(x_i-x)^2}}$

prediction interval $y = \hat{y} \quad \pm t_{\alpha/2,n-2}\sqrt{MSE}\sqrt{1 + \frac{1}{n} + \frac{(x-x)^2}{\sum(x_i-x)^2}}$

$t - statistic\ with\ n-2\ degrees\ of\ freedom\ at\ the\ \frac{a}{2}\ quantile$

a-percentile for two-sided hypothesis test with using the t-distribution

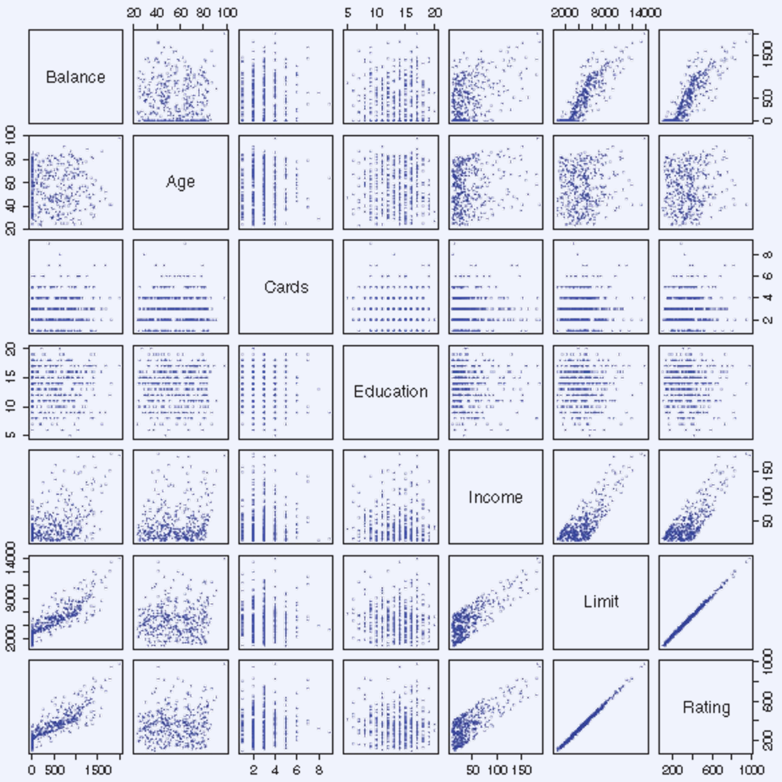
Beyond Simple and Additive

ISLR 3.3

How to Include Qualitative Predictors

Example Credit dataset ($n = 400$)

- output **balance**
- quantitative predictors
  - **age** in years
  - **cards** # credit cards
  - **education** years of education
  - **income** annual, in K$
  - **limit** credit card limit
  - **rating** credit rating
- qualitative predictors
  - **gender** male/female
  - **student** yes/no
  - **status** married/not married
  - **region** 3 values



How to Include Qualitative Predictors

Binary Predictors

- just add a dummy variable, e.g.

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

- which results in a model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

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

- β_0 average credit balance for males
- $\beta_0 + \beta_1$ avg credit balance for females

Alternatively, we can also code as

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

which would give a model

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

where β_0 is the avg credit over all

The choice of coding changes the interpretation of the coefficients but not the regression result

	Coefficient	Std. error	t-statistic	p-value
intercept	509.80	33.13	15.389	<0.0001
gender	19.73	46.05	0.429	0.6690

How to Include Qualitative Predictors

Multiway Predictors (here 3 way)

- use multiple dummy variables

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person from the South} \\ 0 & \text{if } i\text{th person is not from the South} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is from the West} \\ 0 & \text{if } i\text{th person is not from the West} \end{cases}$$

- which results in a model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i + \epsilon_i$$
$$= \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is from the South} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is from the West} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is from the East} \end{cases}$$

One dummy less than the number of values

- β_0 avg balance for East (base line)
- $\beta_0 + \beta_1$ avg balance for South
- $\beta_0 + \beta_2$ avg balance for West

Testing significance

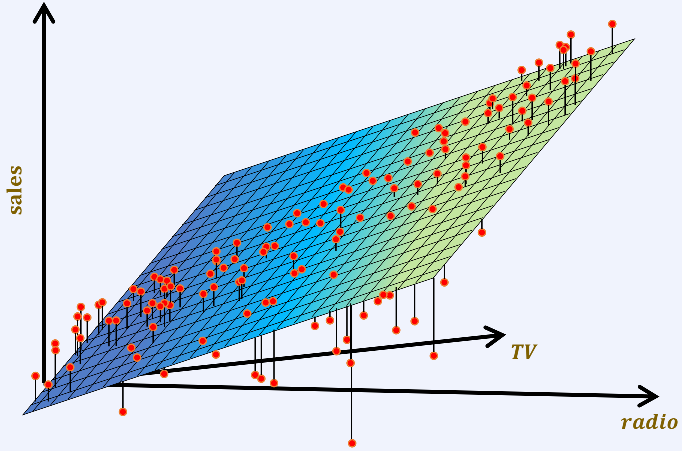
- $H_0: \beta_1 = \beta_2 = 0$, and we use the F-statistic
- we can mix quantitative and qualitative predictors
- we get very high p-values, there is no evidence to reject the null hypothesis

	Coefficient	Std. error	t-statistic	p-value
intercept	531.00	46.32	11.464	<0.0001
region[South]	-18.69	65.02	-0.287	0.7740
region[West]	-12.50	56.68	-0.221	0.8260

How to Account for Interactions among Predictors

Often, additivity does not hold

- e.g. advertising on **radio** can increase the effectiveness of **TV** advertising (synergy)
- the figure shows that the two variables interact
- when levels of either **TV** or **radio** are low then sales are lower than the linear model suggests
- we can account for this by adding an interaction term



For example, we can assume

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$

- where the interaction can be seen as rewriting the model as

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

Example Beyond Additivity

$$\begin{aligned}\textbf{sales} &= \beta_0 + \beta_1\times\textbf{TV} + \beta_2\times\textbf{radio} + \beta_3\times(\textbf{radio}\times\textbf{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3\times\textbf{radio})\times\textbf{TV} + \beta_2\times\textbf{radio} + \epsilon\end{aligned}$$

Strong evidence for $H_a: \beta_3 \neq 0$

- β_3: increase in effectiveness of **TV** advertising per unit increase in **radio** advertising
- $R^2 = 89.7\%$ for the model without the interaction term
- $R^2 = 96.8\%$ for the model with the interaction term
- $(96.8 - 89.7)/(100 - 89.7) = 69\%$ of the unexplained variability is explained by the interaction term
- all terms are significant

	Coefficient	Std. error	t-statistic	p-value
intercept	6.7502	0.248	27.23	<0.0001
TV	0.0191	0.002	12.70	<0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	<0.0001

Accounting for Mixed-Type Interactions

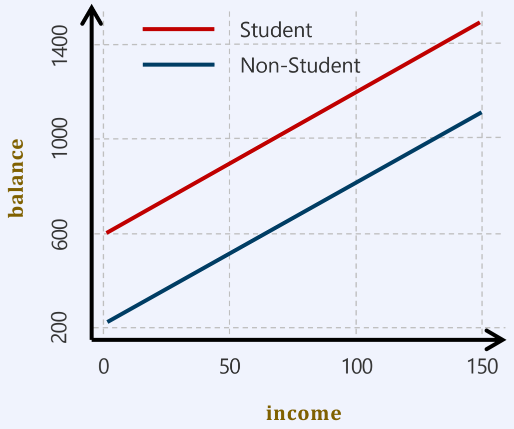
Example credit data with output **balance** and
inputs **income** (quantitative) and **student** (qualitative)

Base model

$$\mathbf{balance}_i = \beta_0 + \beta_1 \times \mathbf{income}_i + \begin{cases} \beta_2 & \text{if the } i\text{th person is a student} \\ 0 & \text{if the } i\text{th person is not a student} \end{cases}$$

$$= \beta_1 \times \mathbf{income}_i + \begin{cases} \beta_0 + \beta_2 & \text{if the } i\text{th person is a student} \\ \beta_0 & \text{if the } i\text{th person is not a student} \end{cases}$$

- forms two parallel lines



Accounting for Mixed-Type Interactions

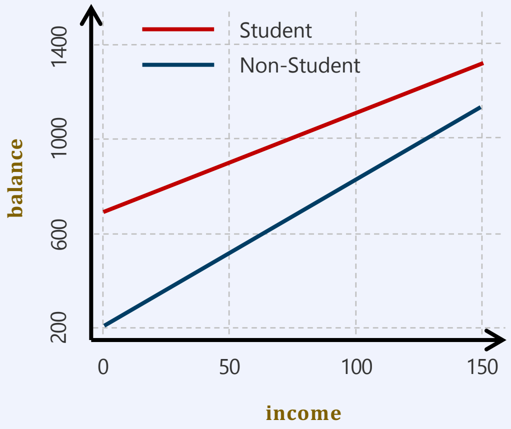
Example credit data with output **balance** and inputs **income** (quantitative) and **student** (qualitative)

Interaction model

$$\textbf{balance}_i = \beta_0 + \beta_1 \times \textbf{income}_i + \begin{cases} \beta_2 + \beta_3 \times \textbf{income}_i & \text{if a student} \\ 0 & \text{if not a student} \end{cases}$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \textbf{income}_i & \text{if a student} \\ \beta_0 + \beta_1 \times \textbf{income}_i & \text{if not a student} \end{cases}$$

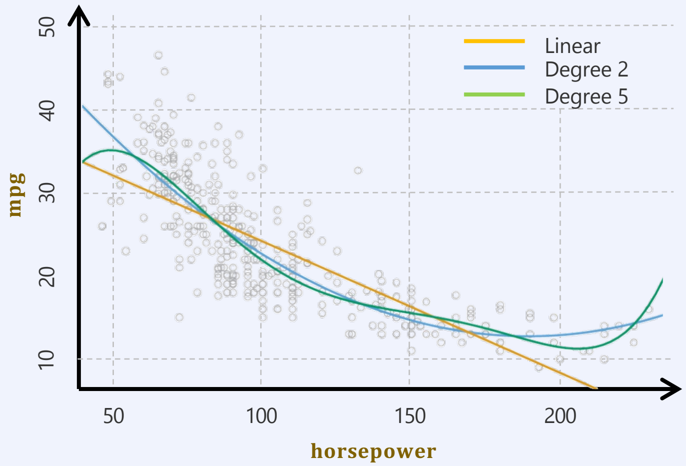
- interaction term allows for different slopes of the two lines



Nonlinear Relationships

We can include non-linearities into linear regression by considering nonlinear functions of the inputs as features

- the functions over inputs are called base functions
- in polynomial regression we consider polynomials over the inputs as base functions, e.g. X_i^2 or X_i^{42}



Example Mileage dataset

- Output, miles per gallon of gas (**mpg**)
- 397 samples, here we consider input **horsepower**

$$\textbf{mpg} = \beta_0 + \beta_1 \times \textbf{horsepower} + \beta_2 \times \textbf{horsepower}^2 + \epsilon$$

	Coefficient	Std. error	t-statistic	p-value
intercept	56.9001	1.8004	31.6	<0.0001
horsepower	-0.4662	0.0311	-15.0	<0.0001
horsepower^2	0.0012	0.0001	10.1	<0.0001

	R^2
linear	0.606
quadratic	0.688

Polynomial regression of degree 5 overfits

Regression Pitfalls

ISLR 3.3.3

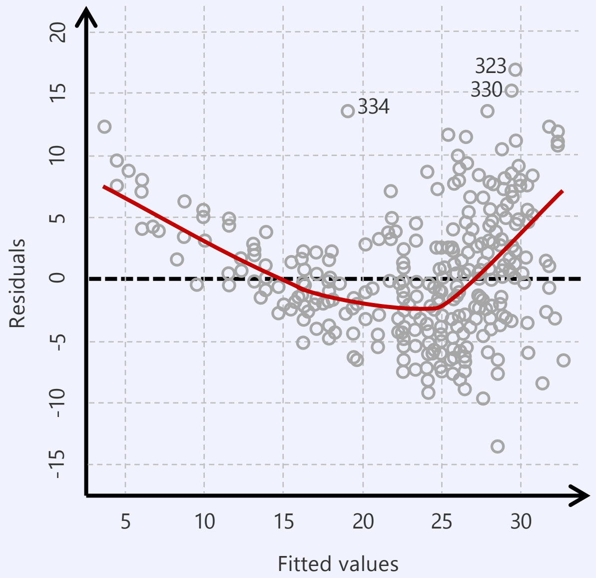
Problem 1. Nonlinearity

If the true relationship is nonlinear, any linear model will be inexact and lead to wrong interpretations

- residual plots can help identify nonlinearity

Plot residual error against the fitted output value

- linear model: U-shape is indicative of non-linear relationship



Residual plot for linear fit of
$\mathbf{mpg} = \beta_0 + \beta_1 \times \mathbf{horsepower} + \epsilon$

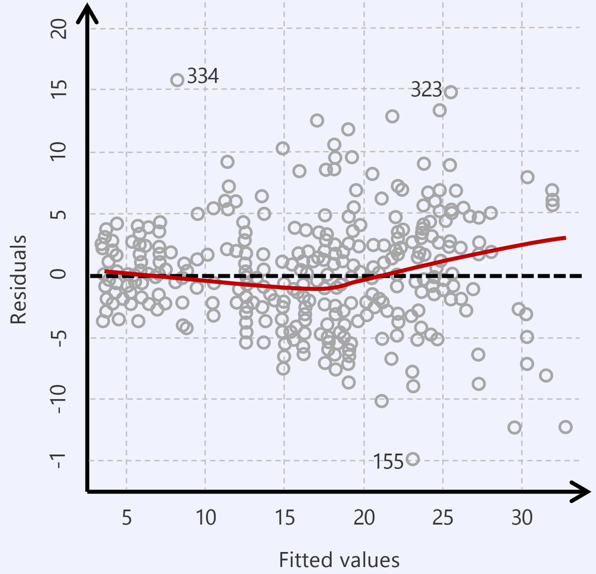
Problem 1. Nonlinearity

If the true relationship is nonlinear, any linear model will be inexact and lead to wrong interpretations

- residual plots can help identify nonlinearity

Plot residual error against the fitted output value

- linear model: U-shape is indicative of non-linear relationship
- quadratic model: curve is flatter, fits the data better
- not perfect, perhaps we should try other base functions…
- Chapter 7 details nonlinear models



Residual plot for quadratic fit of
$\mathbf{mpg} = \beta_0 + \beta_1 \times \mathbf{horsepower} + \beta_2 \times \mathbf{horsepower}^2 + \epsilon$

Problem 2. Correlation of Error Terms

The theory of linear models assumes that the errors $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ are uncorrelated

- if they are correlated, standard errors will be larger than given by the formulas
- confidence and prediction intervals should then be wider and p-values should be higher
- parameters that seem statistically significant, may not be

For example, assume we duplicate our data

- ignoring correlation of errors, we now have a sample of size $2n$
- our coefficients would be the same, but our confidence intervals are narrower by a factor $\sqrt{2}$

$$SE(\hat{\mu}) = \sqrt{Var(\hat{\mu})} = \sqrt{\sigma^2/n}$$

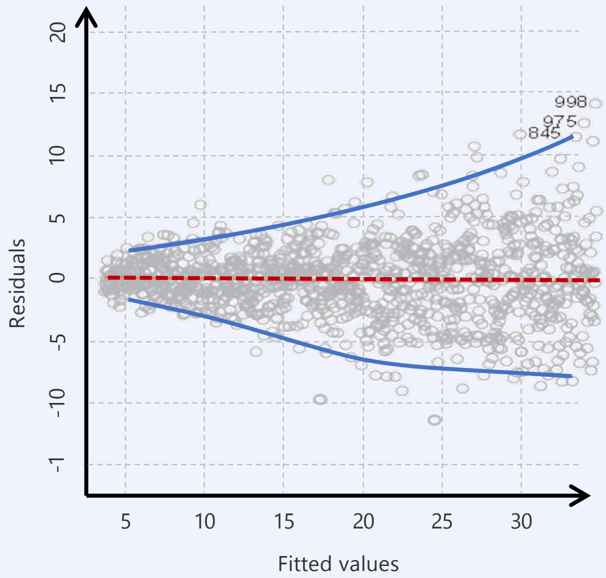
Errors are frequently (positively) correlated

- e.g. adjacent time points in temporal data
- always check for correlated errors, by correlation analysis or by plotting them

Problem 3. Heteroscedasticity

Requiring that $Var(\epsilon_i) = \sigma^2$ is constant is another central assumption in the theory on linear models

- often the variance of the error **depends** on the response
- changing variance is called **heteroscedasticity**
- can be seen as a funnel shape in the residual plot



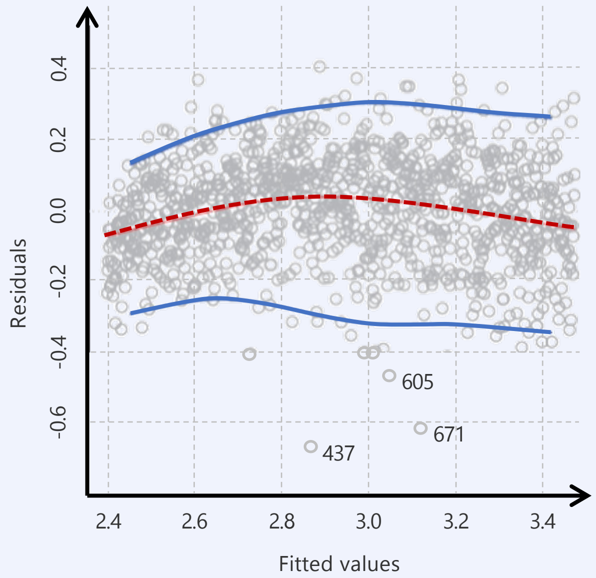
Response Y
Red line is the moving average
Blue lines delineate outer quantiles of the plot

Problem 3. Heteroscedasticity

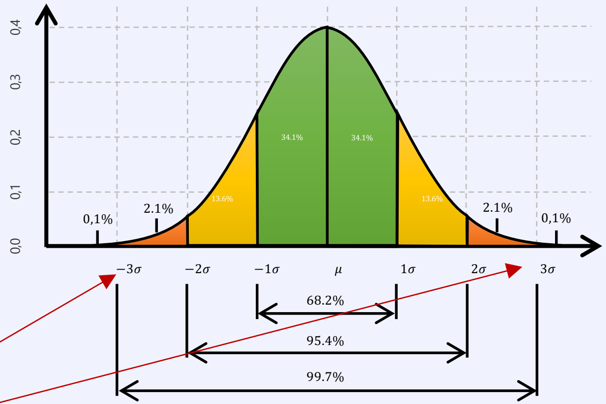
Requiring that $Var(\epsilon_i) = \sigma^2$ is constant is another central assumption in the theory on linear models

- often the variance of the error **depends** on the response
- changing variance is called **heteroscedasticity**
- can be seen as a non-uniform shape in the residual plot
- can (often) be dealt with by transforming the response using a concave function

If we know how the variance depends on the response we can weigh observations to even out the variance

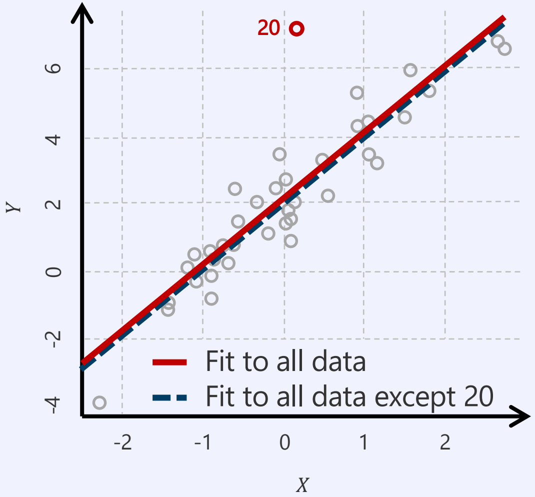


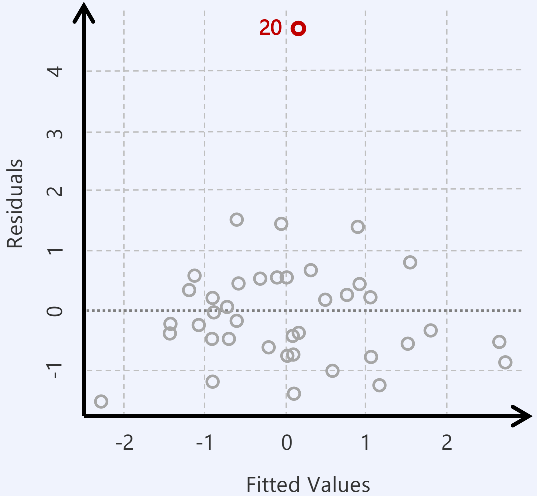
Problem 4. Outliers

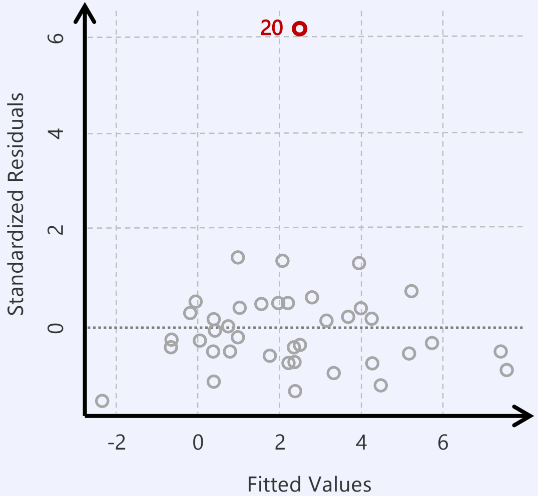


Outliers are data points whose outcome is far from prediction

- residuals can identify outliers
- when is a residual large enough to call a point an outlier?
- **studentized residuals:** divide residuals by its estimated standard error
- if absolute studentized residual is >3 a point is an outlier



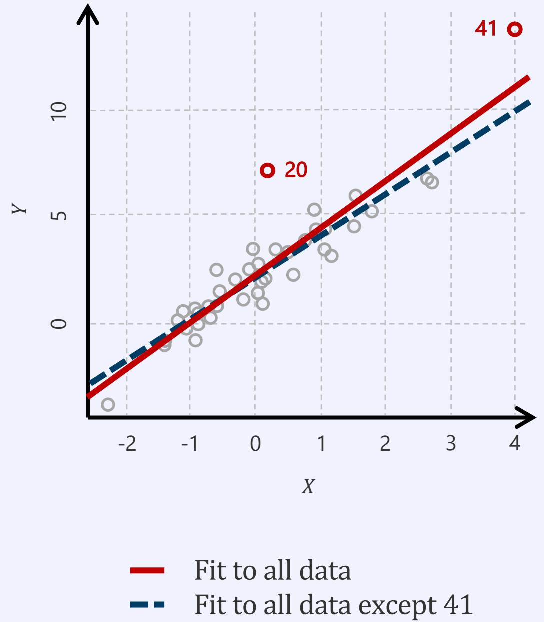




Problem 5. High Leverage Points

Data points with unusual (unlikely) input values x_i

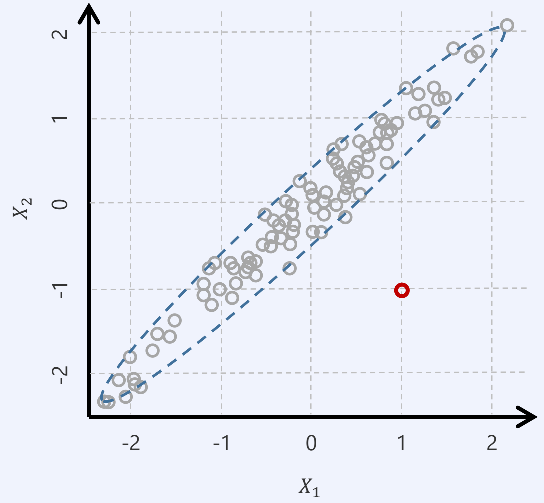
- for example, point 41 in the figure
- high leverage points have large impact on the regression line
- important to identify (and potentially remove) these points



Problem 5. High Leverage Points

Data points with unusual (unlikely) input values $\boldsymbol{x_i}$

- for example, point 41 in the figure
- high leverage points have large impact on the regression line
- important to identify (and potentially remove) these points



Identifying high leverage points is difficult in high-dimensions

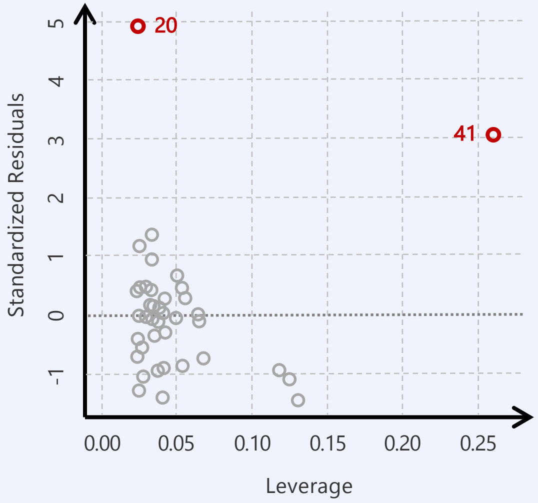
- thus we compute and use the leverage statistic
- for univariate data

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

Problem 5. High Leverage Points

Points with unusual (unlikely) input values x_i

- for example, point 41 in the figure
- high leverage points have large impact on the regression line
- important to identify (and potentially remove) these points



Identifying high leverage points is difficult in high-dimensions

- thus we compute and use the leverage statistic
- for univariate data

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

- for multivariate data
  h_{ii} is the ith diagonal element of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, which effectively tells us the influence of y_i on $\hat{y}_i$

Problem 6. Collinearity

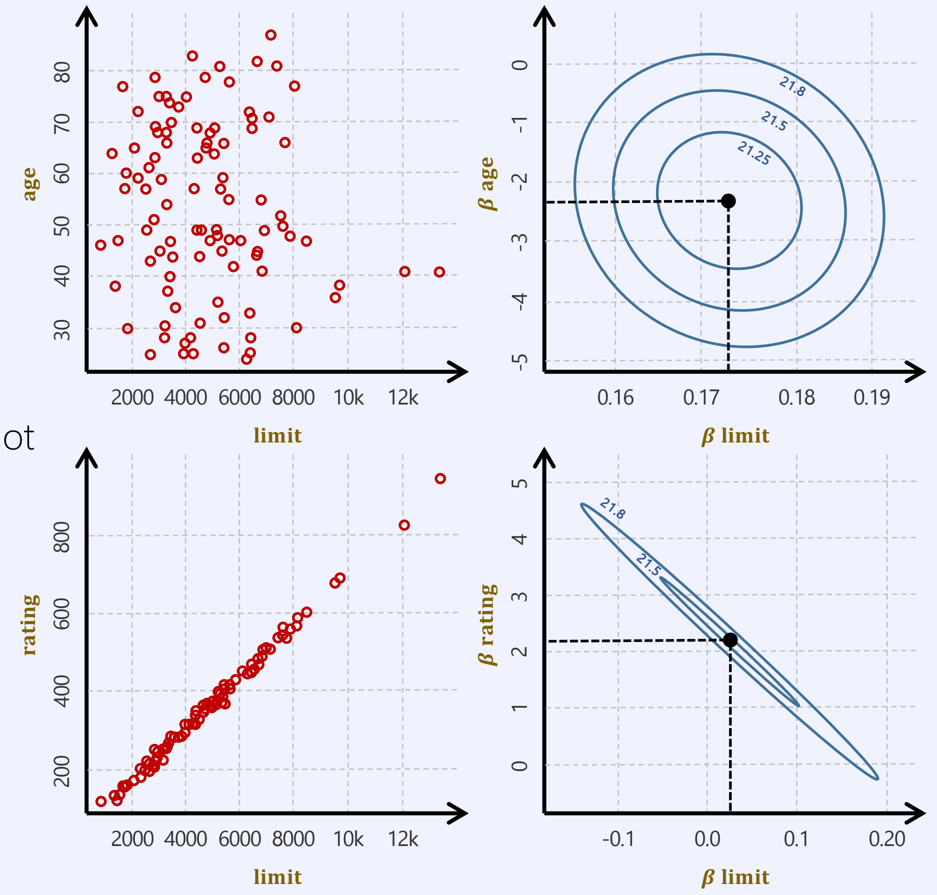
Two related predictors are called collinear

- they can substitute for each other
- i.e. trade parts of their coefficients
- results in large variance in the model

In the credit data

- **rating** and **limit** are collinear; **age** and **limit** are not

Model 1		Coefficient	Std. error	t-statistic	p-value
	intercept	-173.411	43.828	-3.957	<0.0001
	age	-2.292	0.672	-3.407	0.0007
	limit	0.173	0.005	34.496	<0.0001
Model 2		**Coefficient**	**Std. error**	**t-statistic**	**p-value**
	intercept	-377.537	45.254	-8.343	<0.0001
	rating	2.202	0.952	2.312	0.0213
	limit	0.025	0.064	0.384	0.7012



Problem 6. Collinearity

We can detect pairwise collinearity by looking at the correlation matrix of the predictors

- collinearity among larger sets of predictors (multi-colinearity) cannot be seen this way!
- the variance of a coefficient decomposes as

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{(n-1)Var(X_j)} VIF(\hat{\beta}_j)$$

- where VIF stands for the variance inflation factor

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

R^2 of regressing X_j on all other X_i

- $VIF = 1$ if there is no collinearity, larger otherwise, where a $VIF \geq 5$ or $VIF \geq 10$ indicates a problem

How to handle collinearity

1. drop problematic variable from the data
   - in the example, dropping **rating** reduces all VIFs to ≈ 1 while R^2 drops only from 0.754 to 0.75
2. combine the collinear variables into a single predictor, e.g. by averaging

	VIF
age	1.01
rating	160.67
limit	160.59

(the book is sparse on this topic, see https://en.wikipedia.org/wiki/Variance_inflation_factor for more details)

kNN vs. Linear Regression

ISLR 3.5

k-NN Regression

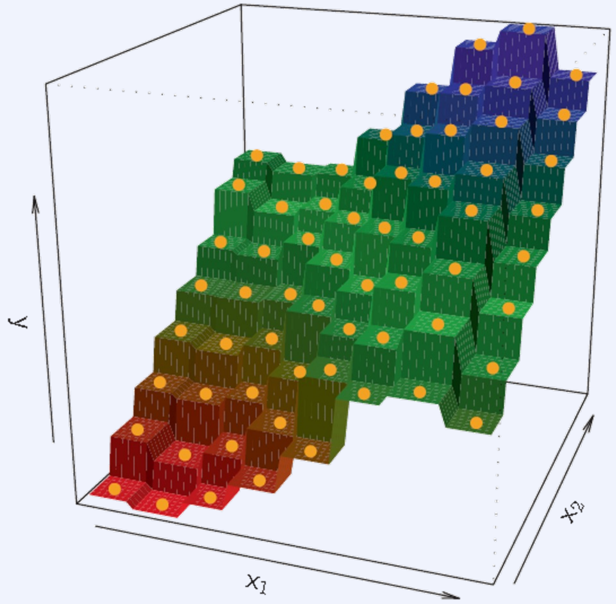
k-NN regression

$$\hat{f}(x_0) = \frac{1}{k} \sum_{x_i \in \mathcal{N}_0} y_i$$

- optimal value of k depends on the bias-variance tradeoff

Small values of k leads to complex models

- high variance: single point can strongly affect the model



$k = 1$

k-NN Regression

k-NN regression

$$\hat{f}(x_0) = \frac{1}{k} \sum_{x_i \in \mathcal{N}_0} y_i$$

- optimal value of k depends on the bias-variance tradeoff

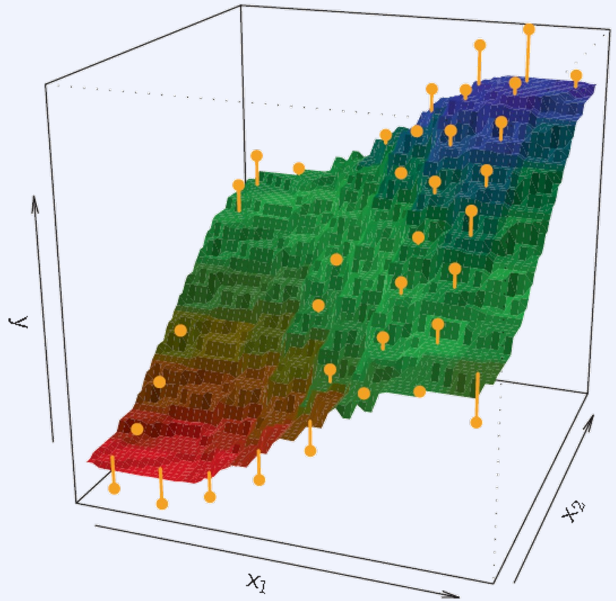
Small values of k leads to complex models (likely to overfit)

- high variance: single point can strongly affect the model

Large values of k leads to simple models (likely to underfit)

- high bias: model becomes too smooth

Optimal value of k can be found by estimating the test error (Ch. 5)



$k = 9$

Comparing kNN and Linear Models

Which model should we use?

- the one that mimics the data (reality!) best

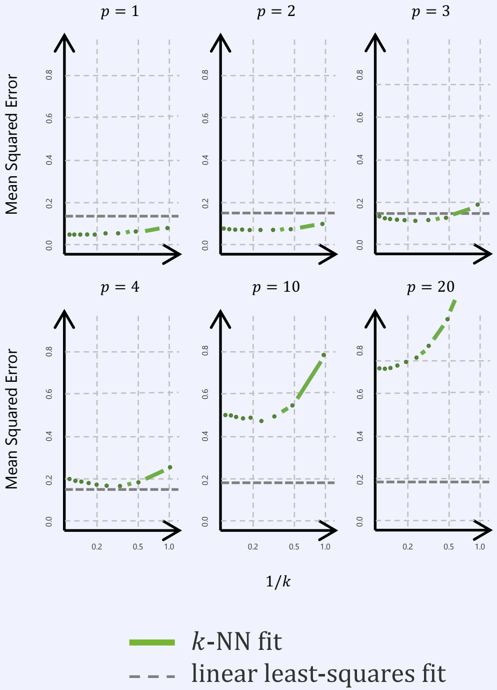
Linear Models

- assume the whole world is linear (parametric)
- linear models are easily interpretable and provide p-values

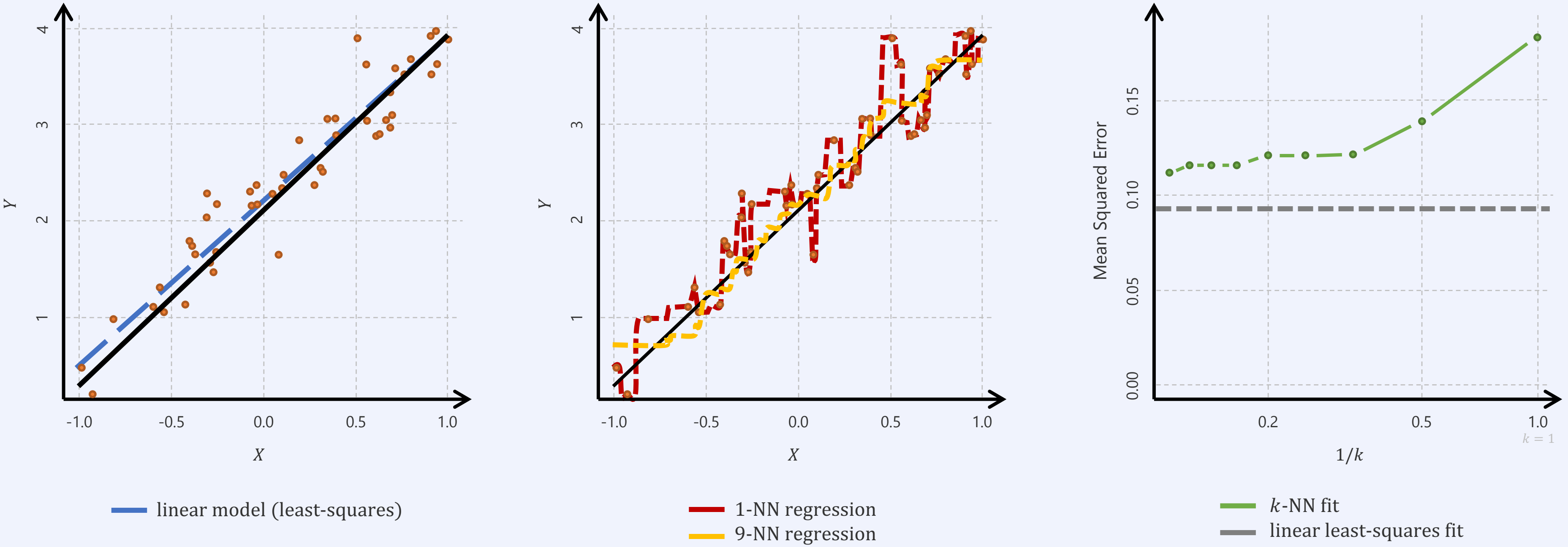
k-NN Models

- assume the world is locally constant
- non-parametric, at least for small k
- adding noise variables upsets k-NN more than linear models
- in high dimensions every point is far away (curse of dimensionality)

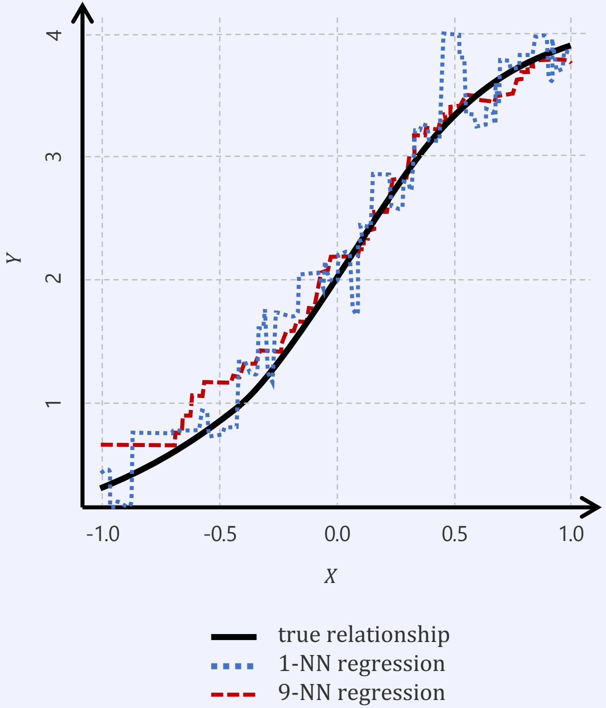
True function strongly nonlinear in first variable, independent of all other variables

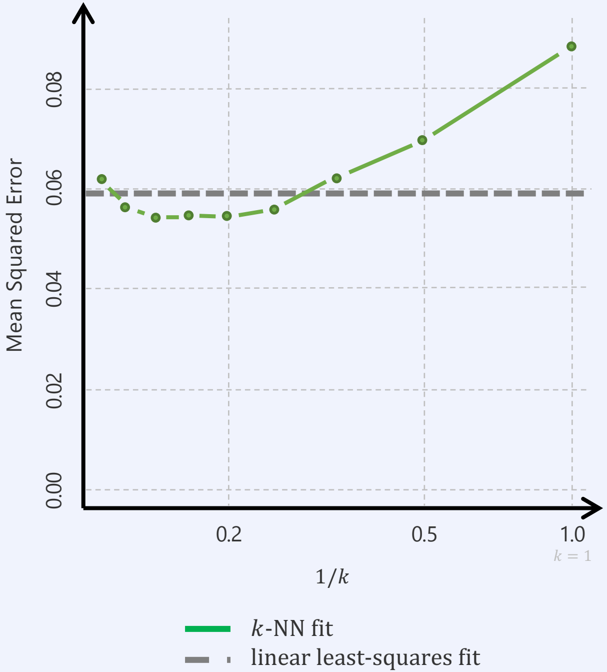


*k*NN vs. Linear on Linear Data



*k*NN vs. Linear on Mildly Non-Linear Data





***k*NN vs. Linear on Non-Linear Data**