

Data Mining (Week 1)

dm25s1

Introduction

Topic 09 : Regression2

Part 01: Overview

Preparation

Data Handling

Dr Bernard Butler

Department of Computing and Mathematics, WIT.

Exploring Data (bernard.butler@setu.ie)

Exploring Data 2

Building Models

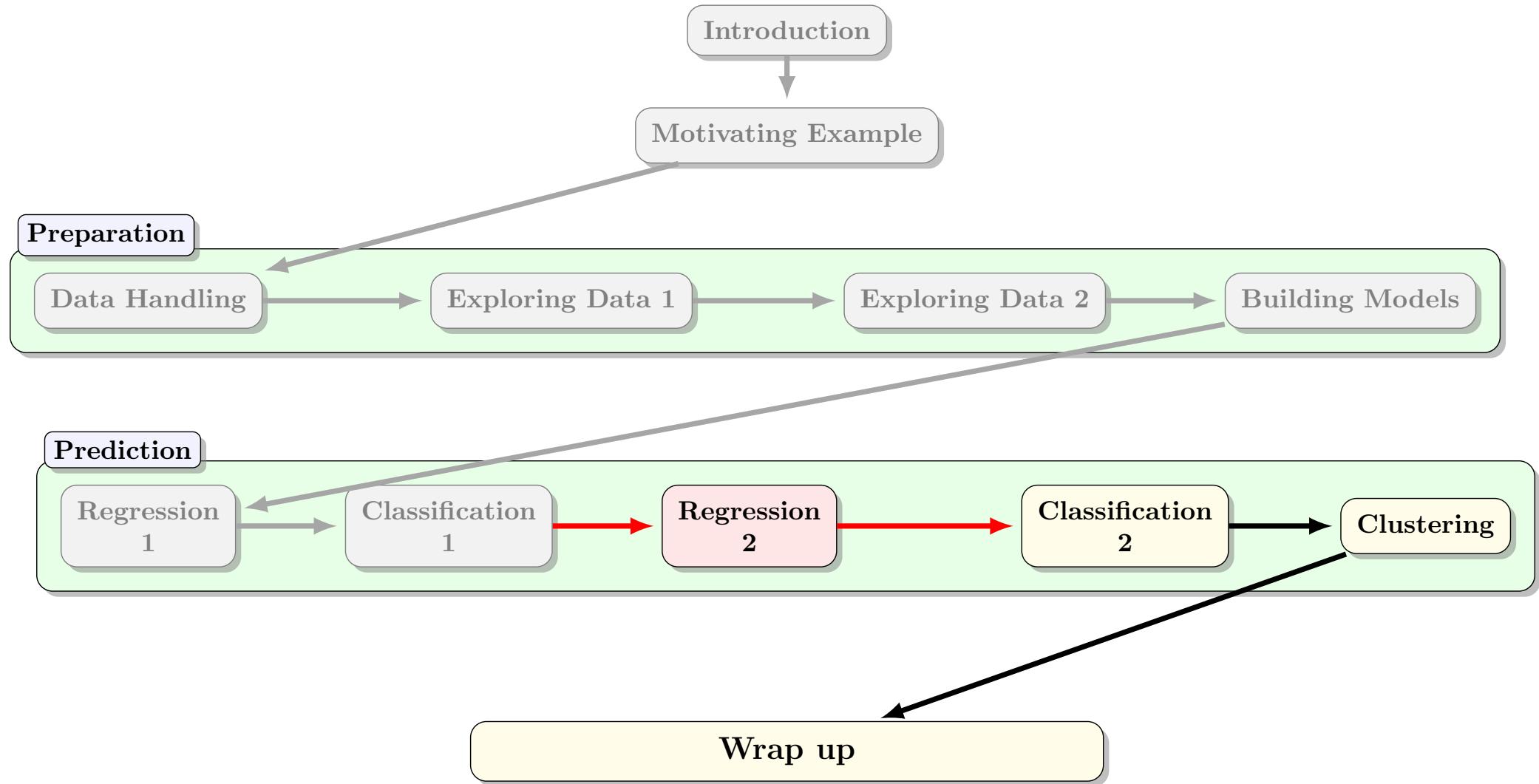
Autumn Semester, 2025

Outline

- Regression assumptions, and how-
- to deal with heteroscedasticity and why it is a problem
- unrepresentative training data can lead to overfitting
- feature collinearity can be assessed
- Provide a worked example of forward selection of features, and interaction terms, for model building

Wrap up

Data Mining (Week 9)



Overview — Summary

1. Introduction	4
2. Regression1 review	6
3. Case Study 1: Generated	10
4. Case Study 3: Advertising	15

This Week's Aim

This week's aim is to continue the introduction to linear regression, focusing more on how to deal with problems with more challenging datasets.

- Examine some extensions to the simplest case of linear regression.
- We introduce two new concepts: dimensionality reduction and regularisation
- To provide context we will use the following datasets:
 - Generated data (various)
 - Advertising dataset: predicting widgets sold based on spending in different advertising channels
 - Credit dataset: predicting credit balance using income, status, etc.

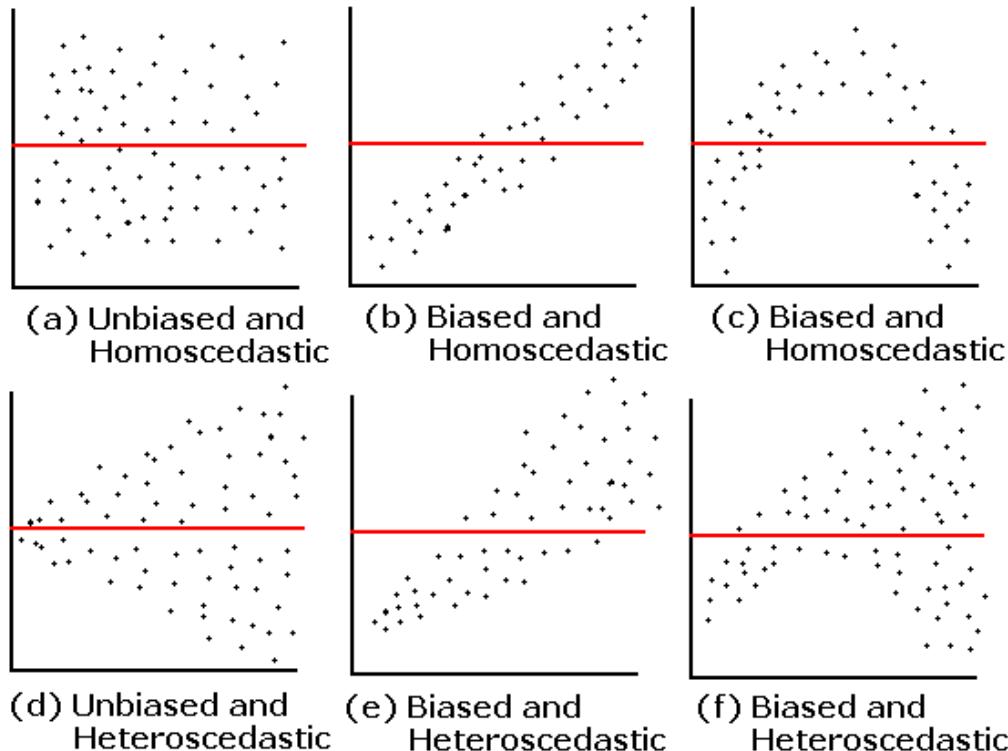
Assumptions required for the linear model to be meaningful

Definition 1 (Linear Regression Assumptions)

- ① The underlying relationship between the predictors and the response is linear in the regression parameters β .
- ② The residual errors ϵ are drawn from a (multivariate) Normal distribution $N(\mu, \sigma^2)$ where $\mu = \mathbf{0}$.
- ③ The predictors are not pairwise collinear, i.e., each pair of predictors β_{j_1} and β_{j_2} (associated with columns $X(:, j_1)$ and $X(:, j_2)$) have low correlation (equivalently, the inner product of $X(:, j_1)$ and $X(:, j_2)$ is far from zero).
- ④ There is no auto-correlation in y : each observation is independent of its “neighbours”.
- ⑤ The errors are *homoscedastic* (i.e., $\text{Var}(\epsilon)$ is constant over the range of x or y).

➤ These assumptions can be used constructively, when model building, or as checks, when validating models.

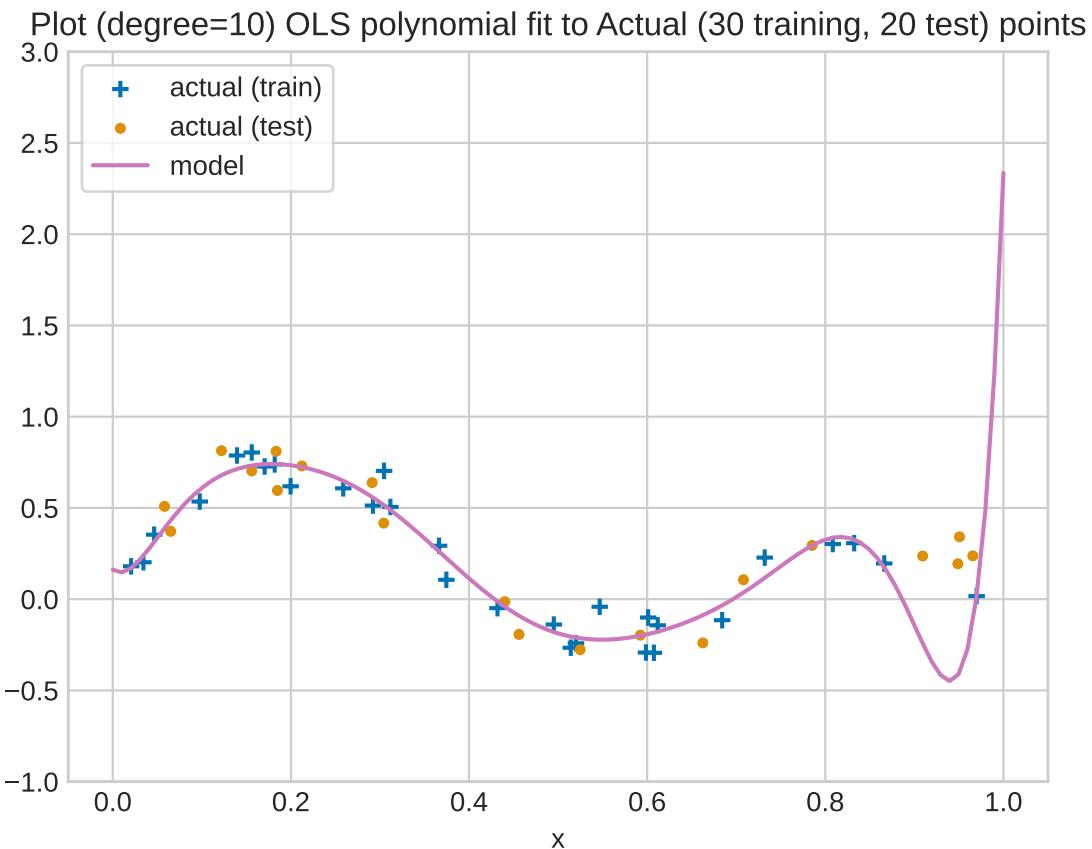
Bias and variance in regression



Source: <https://bit.ly/3vC9zK7>

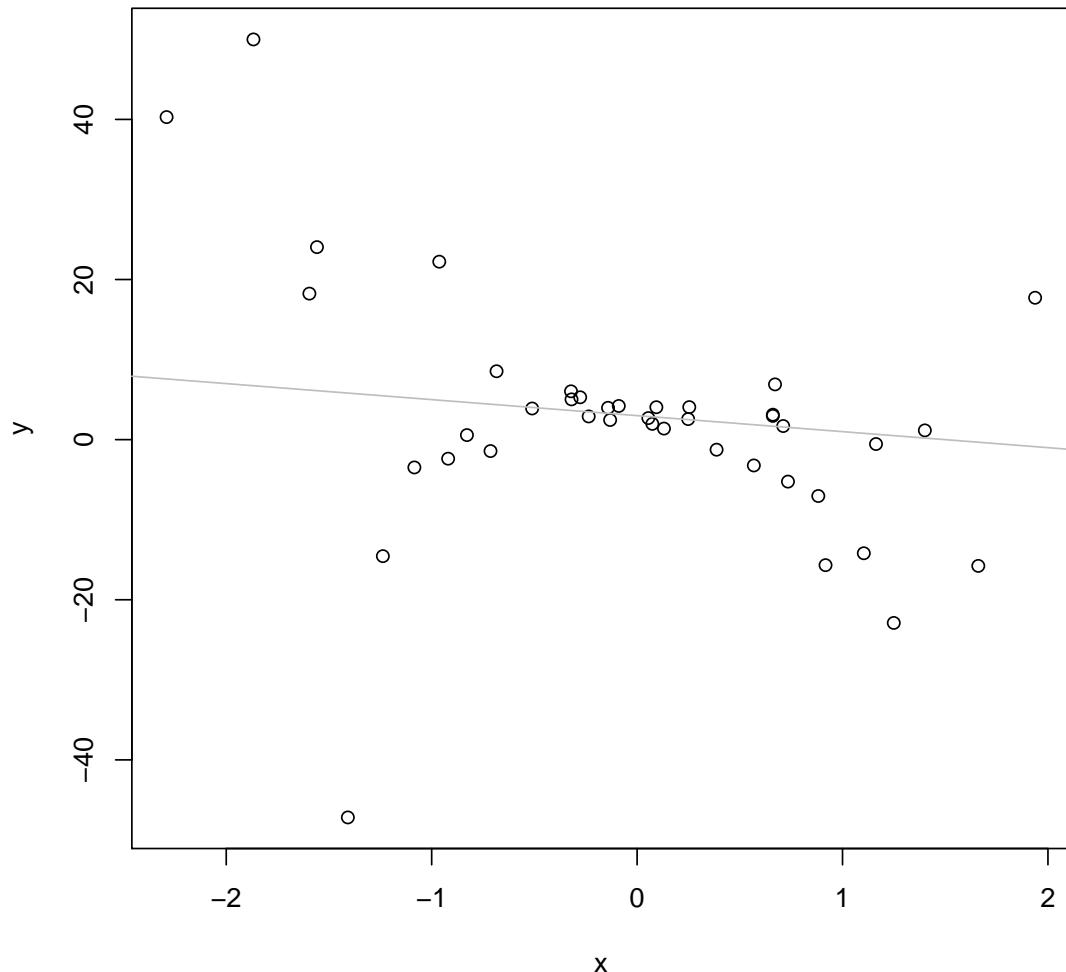
- Bias is caused by underfitting.
- Fix bias by adding suitable predictors.
- Overfitting causes large variance.
- If variance changes over the range, some errors get undue attention.
- Fix this by weighting the errors so the weighted errors satisfy $w_i e_i \approx w_j e_j, \forall i, j$.
- In practice, $w_i \approx \frac{1}{\text{Var}(e_i)}$.
 - Using scikit-learn: add the argument `sample_weight = someWeights`, e.g., `model.fit(Xtrain, yTrain, sample_weight=someWeights)`.
 - Using statsmodels: use the weighted version of least squares: `WLS(y, X, someWeights)` not `OLS(y, X)`

What's happening here???



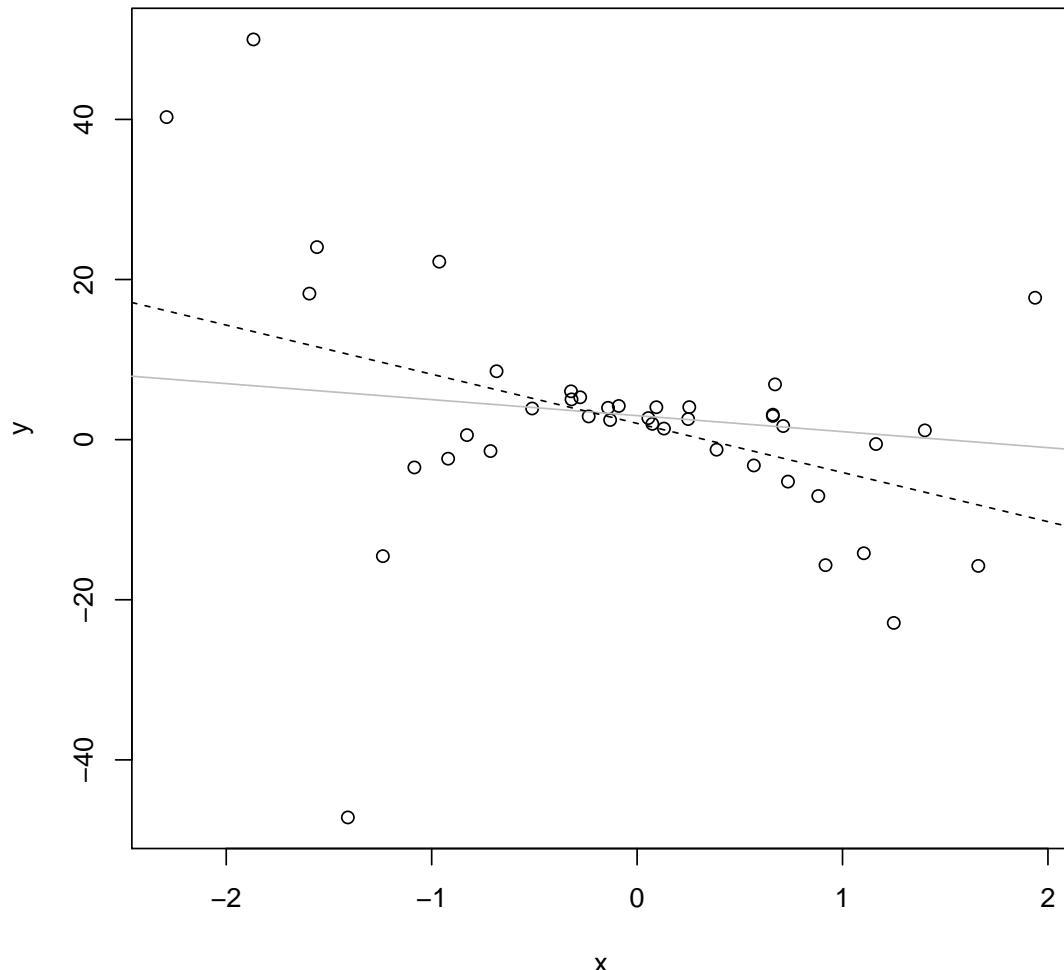
- ➊ Data is quite noisy
- ➋ Training data has gaps near the edges
- ➌ Model may be overfitting

Case Study 1: Heteroscedasticity - Step 1



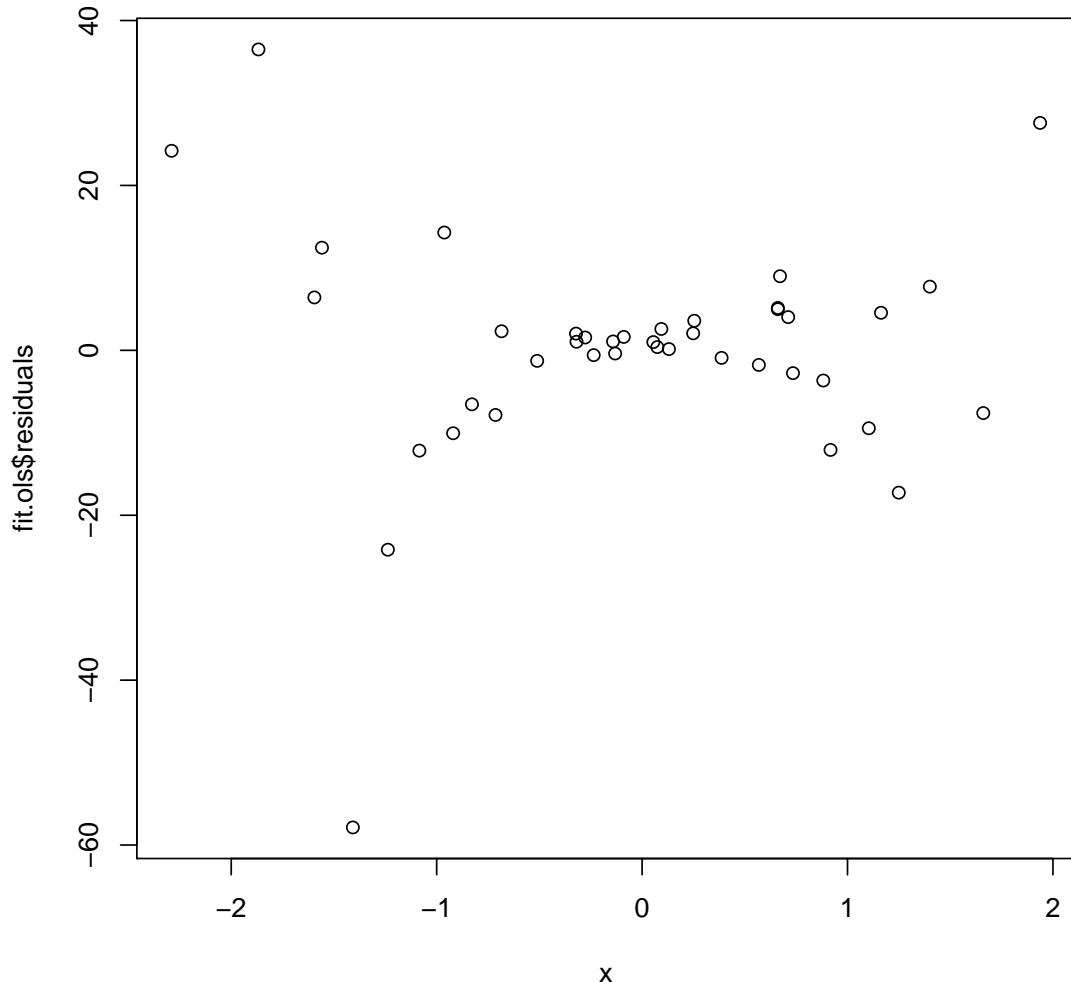
I generated 41 x, y points based on $y = 3 - 2x$, but with added errors that increase away from $x = 0$. The plot shows the line with $\beta = (3, -2)$ in grey.

Case Study 1: Heteroscedasticity - Step 2



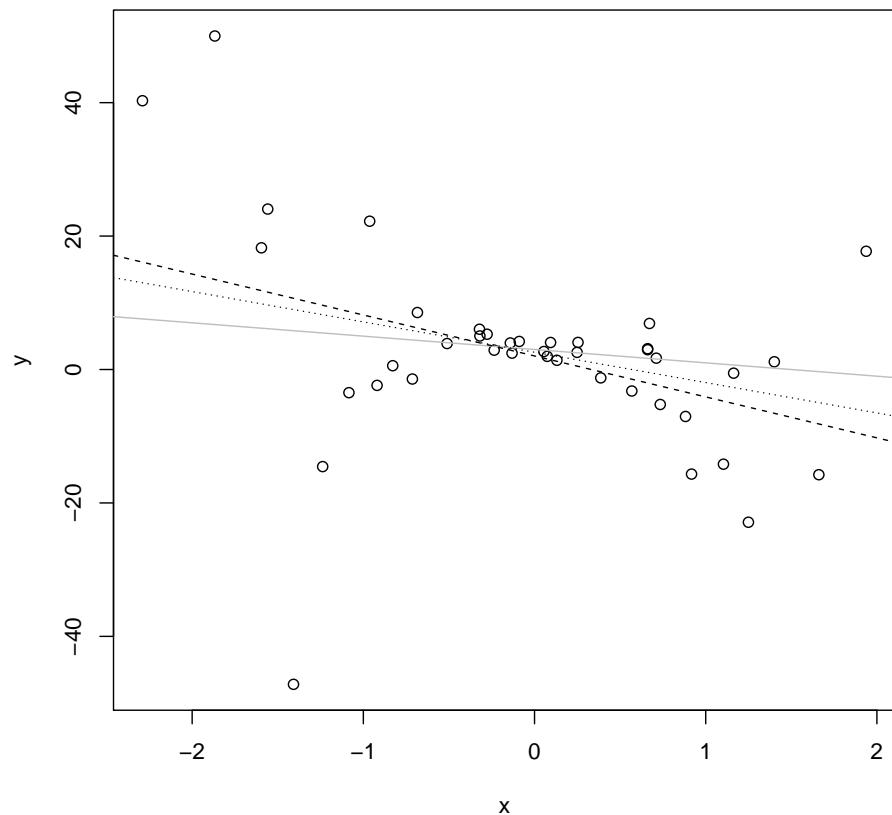
In this plot I added the OLS fit as a dashed line. Note that the parameters of the fit are quite different:
 $\beta_{OLS} \approx (2, -6)$, equivalent to
 $y = 2 - 6x$.

Case Study 1: Heteroscedasticity - Step 3



This plot shows how the OLS residuals ϵ_{OLS} increase rapidly away from 0, as expected (since this was how the data was generated).

Case Study 1: Heteroscedasticity - Step 4



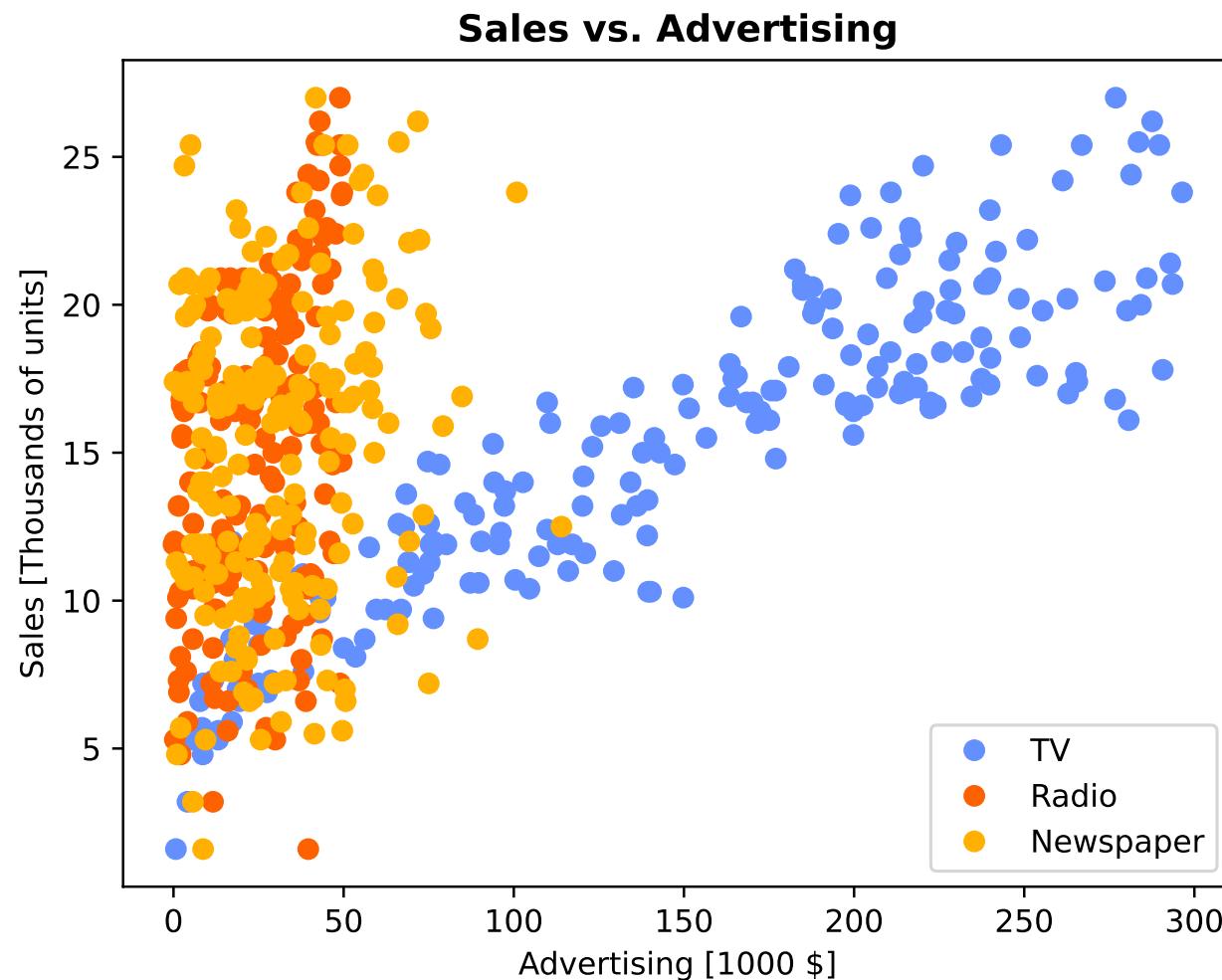
- By inspecting the previous residual plot I estimated a weighting function so that the residuals would be “more constant”. When this was used to scale the residuals, the resulting Weighted Least Squares estimates were $\beta \approx (2.6, -4.5)$ (shown as a dotted line) and hence closer to the “true” $\beta = (3, -2)$.
- So we were only partially successful at stripping away the noise and recovering the original line.
- **Can you see a problem with finding the weights?**
- If the weights are computed from the errors, they depend on the fit, hence on the weights!!
- *Iteratively Reweighted Least Squares* has been proposed to optimise regression models.

Case Study 3: Advertising: Data and Hypotheses

	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	12.0
3	151.5	41.3	58.5	16.5
4	180.8	10.8	58.4	17.9

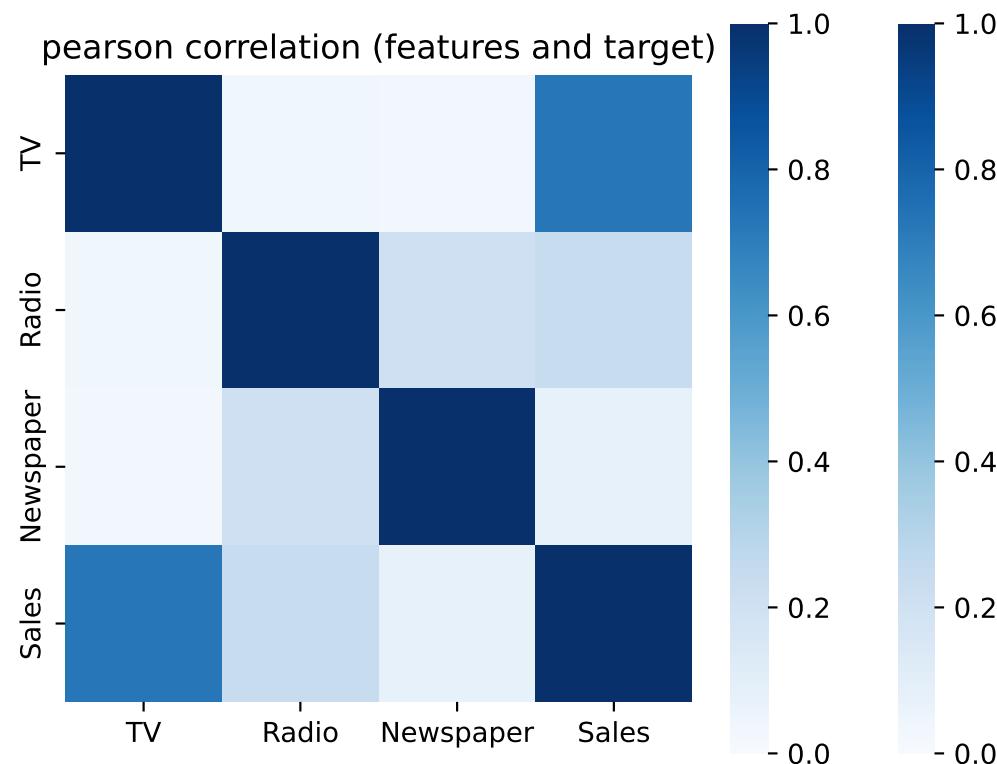
- In this data set, the sales figure captures how many thousands of widgets of a particular type were sold in a year.
- Newspaper, Radio and TV represent the annual spend per widget type on the associated advertising channel.
- The hypothesis is that spend on advertising is a good predictor of sales performance.
- Since marketing budgets are limited, where should the adverts be placed for maximum sales?
- Alternatively, how should marketing funds be distributed across the 3 channels to achieve a specified sales performance, while keeping the total spend as low as possible?

Case Study 3: Advertising: Looking at the data



Which of the advertising channels appear to have a linear relationship with Sales?

Case Study 3: Advertising: Collinearity?



- Correlation matrix can indicate which features should participate in the model as predictors.
- A good predictor should have a **high correlation with the target** (Sales in this case) and should have **low correlation with other candidate predictors**.
- **What are expected to be good predictors for this data?**
 - Sales (the target) is placed in the last row (or column).
 - TV > Radio > Newspaper, with moderate correlation between Radio and Newspaper.

Sidebar: specifying models

The statsmodels way

- The dataframe contains the observed variables
- The model is specified separately
- Easier to change the model when experimenting
- statsmodels models are expressed like "Sales ~ TV * Radio + poly(Newspaper,2)". This notation came from the applied statistics community.
- In words: "Sales depends on TV spending, Radio spending, the interaction between TV and Radio spending, Newspaper spending and Newspaper spending squared (5 features from 3 measured features)."
- statsmodels offers its own plotting (like seaborn but not as good). Its model summary is very convenient.
- sklearn exposes more of the details (e.g., choice of algorithm and configuration parameters).
- Both statsmodels and sklearn use the same libraries (scipy, numpy, etc.) underneath.

The sklearn way

- The dataframe contains the (computed) features
- The model is defined implicitly
- Standard interface across all sklearn

Case Study 3: Advertising: Model Building (“stats” way)

- Start from a “full model” and prune, versus from an “empty model” and add
- We choose the latter, as it is often easier to avoid overfitting

Example 2 (Forward Selection for Advertising Data)

Define: model score: mean-square-error on the test set for a given model.

- ➊ Fit “Sales ~ Newspaper”, “Sales ~ Radio”, “Sales ~ TV” and calculate their loss values.
- ➋ Choose the best (lowest loss) single-term model (“Sales ~ TV” in this case), with loss $MSE(TV)$.
- ➌ Fit “Sales ~ TV + Newspaper” and “Sales ~ TV + Radio” and choose the lowest loss score, which is “Sales ~ TV + Radio” with loss being $MSE(TV + Radio)$, which is significantly better.
- ➍ Fit “Sales ~ TV + Radio + Newspaper”. Its loss is the same ($MSE(TV + Radio) \approx MSE(TV + Radio + Newspaper)$), so we favour the existing simpler two-term model (Occam’s Razor: other things being equal, choose the simplest model.).

So our preferred model is “Sales ~ TV + Radio”.

Forward selection in action, with and without the interaction term

Main features only

feature	test_neg_mean_squared_error	test_r2
0 TV	(-7.324310374422007, -3.936981032219174)	(0.7603440777107349, 0.8390841989031752)
1 Radio	(-4.718440611471559, -1.8510139478354652)	(0.8456097326980662, 0.9322678692463671)
2 Newspaper	(-4.72039259225367, -1.8510521207093062)	(0.8455458626911012, 0.9317779087301497)

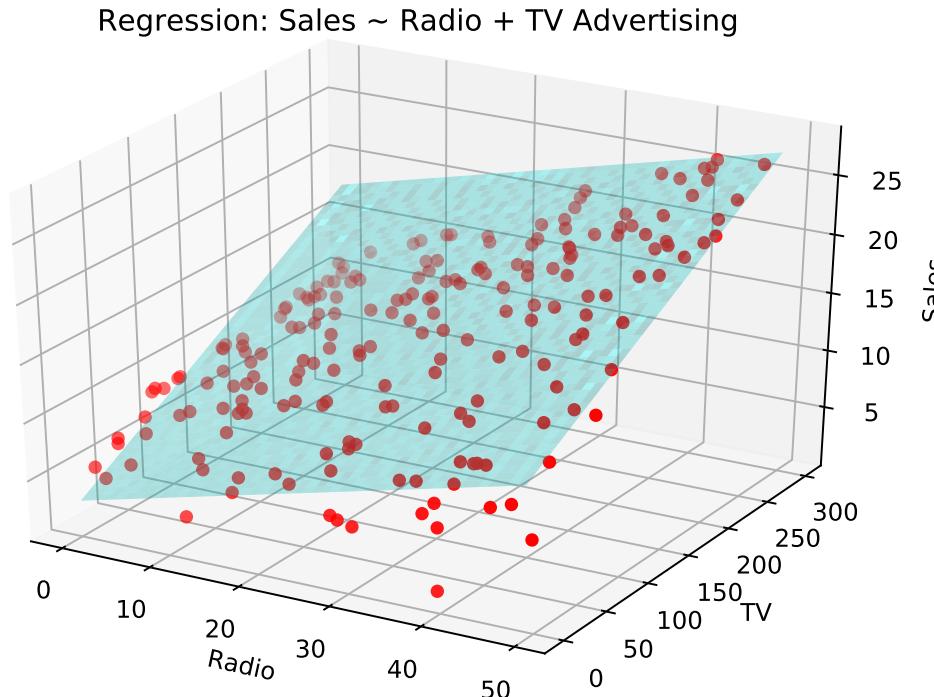
$\text{MSE}(\text{TV}) \approx 5.5$; $\text{MSE}(\text{TV} + \text{Radio}) \approx 3.5$; $\text{MSE}(\text{TV} + \text{Radio} + \text{Newspaper}) \approx 3.5 \approx \text{MSE}(\text{TV} + \text{Radio})$. Adding Newspaper does not reduce MSE.

Main features with TV:Radio interaction term

feature	test_neg_mean_squared_error	test_r2
0 TV	(-7.324310374422007, -3.936981032219174)	(0.7603440777107349, 0.8390841989031752)
1 TV:Radio	(-3.695048288640374, -1.8479935191656154)	(0.8790957564264388, 0.9377953274242408)
2 Radio	(-3.929784758825862, -1.751389612982793)	(0.8714150353235091, 0.9410470781968057)
3 Newspaper	(-3.9387465036567235, -1.7715653928145287)	(0.8711218015427205, 0.9403679482294234)

$\text{MSE}(\text{TV}) \approx 5.5$; $\text{MSE}(\text{TV} + \text{TV:Radio}) \approx 2.8$; $\text{MSE}(\text{TV} + \text{TV:Radio} + \text{Radio}) \approx 2.8 \approx \text{MSE}(\text{TV} + \text{TV:Radio})$. Adding Radio and Newspaper does not reduce MSE.

Case Study 3: Advertising: Viewing the Model



Since this two-term model ignores the contribution of the newspaper channel, the Newspaper spend as a contribution to Sales is just another component of the unmodelled (and apparently random) contribution to Sales.

However, the result is a model where every term is highly significant and the model “explains” 90% of the variance of the data, which is high for an observational study. **Why? Can we do better?**

Case Study 3: Advertising: Interactions; Interpretation

- Trying powers greater than 1 of the Radio and TV features did not offer much more.
- However, by adding the TV, Radio interaction so that the model became “Sales \sim TV + TV:Radio” or equivalently “Sales \sim TV * Radio - Radio”, the loss decreased significantly, indicating the interaction term is valuable, even more so than the Radio feature.
- All β terms have t -statistic significance of approximately 0.001 which is extremely significant.
- $\beta_0 = 6.75$, $\beta_{\text{TV}} = 0.019$, $\beta_{\text{Radio}} = 0.029$ and $\beta_{\text{TV:Radio}} = 0.001$, indicating that there is a favourable relationship between TV and Radio advertising ($\beta_{\text{TV:Radio}} > 0$), and that additional spending on Radio results in more Sales than the same spending on TV ($\beta_{\text{Radio}} > \beta_{\text{TV}}$).
- Spending on Newspaper advertising should be discontinued as its contribution to Sales is insignificant (indistinguishable from random noise).