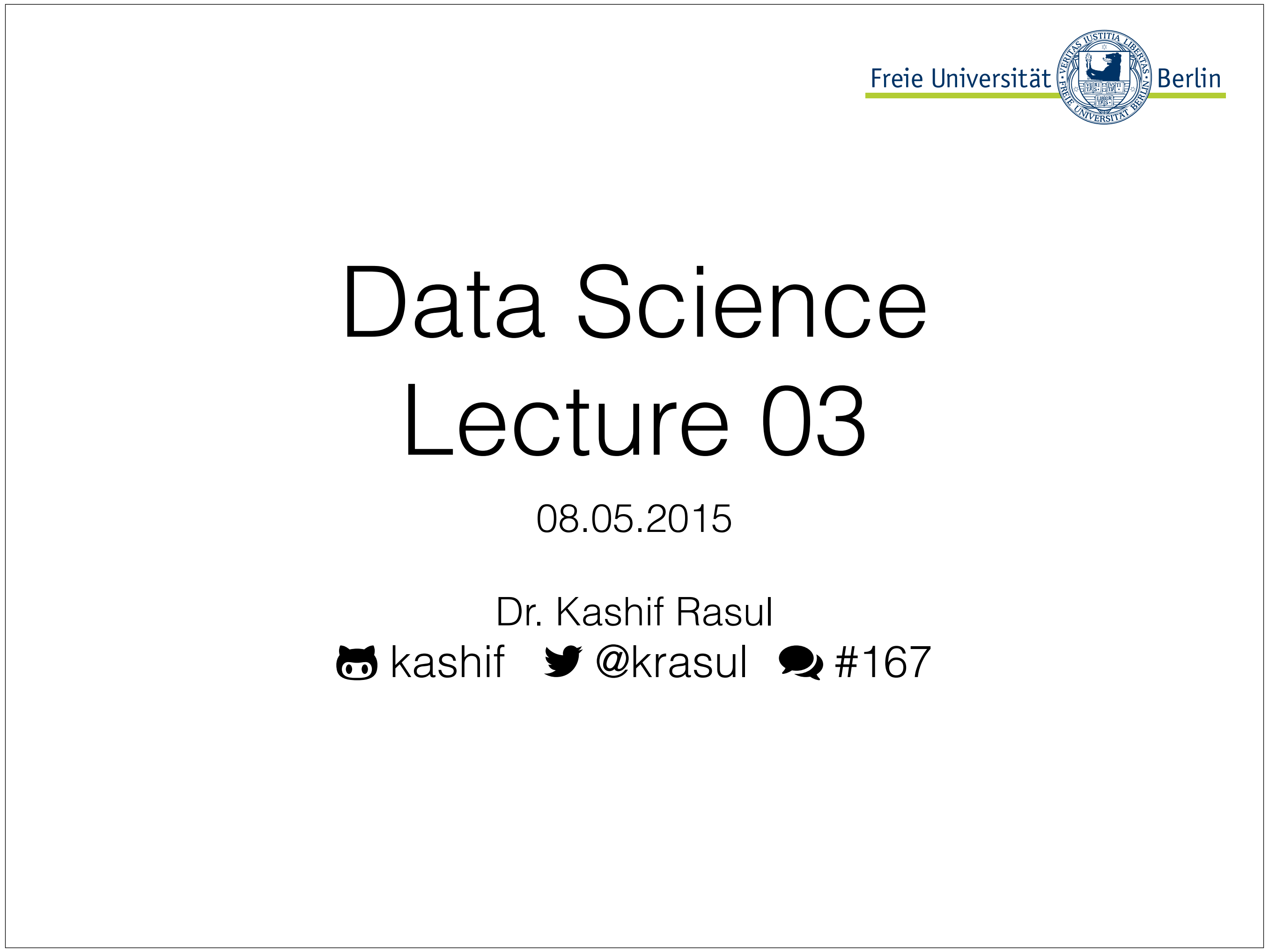


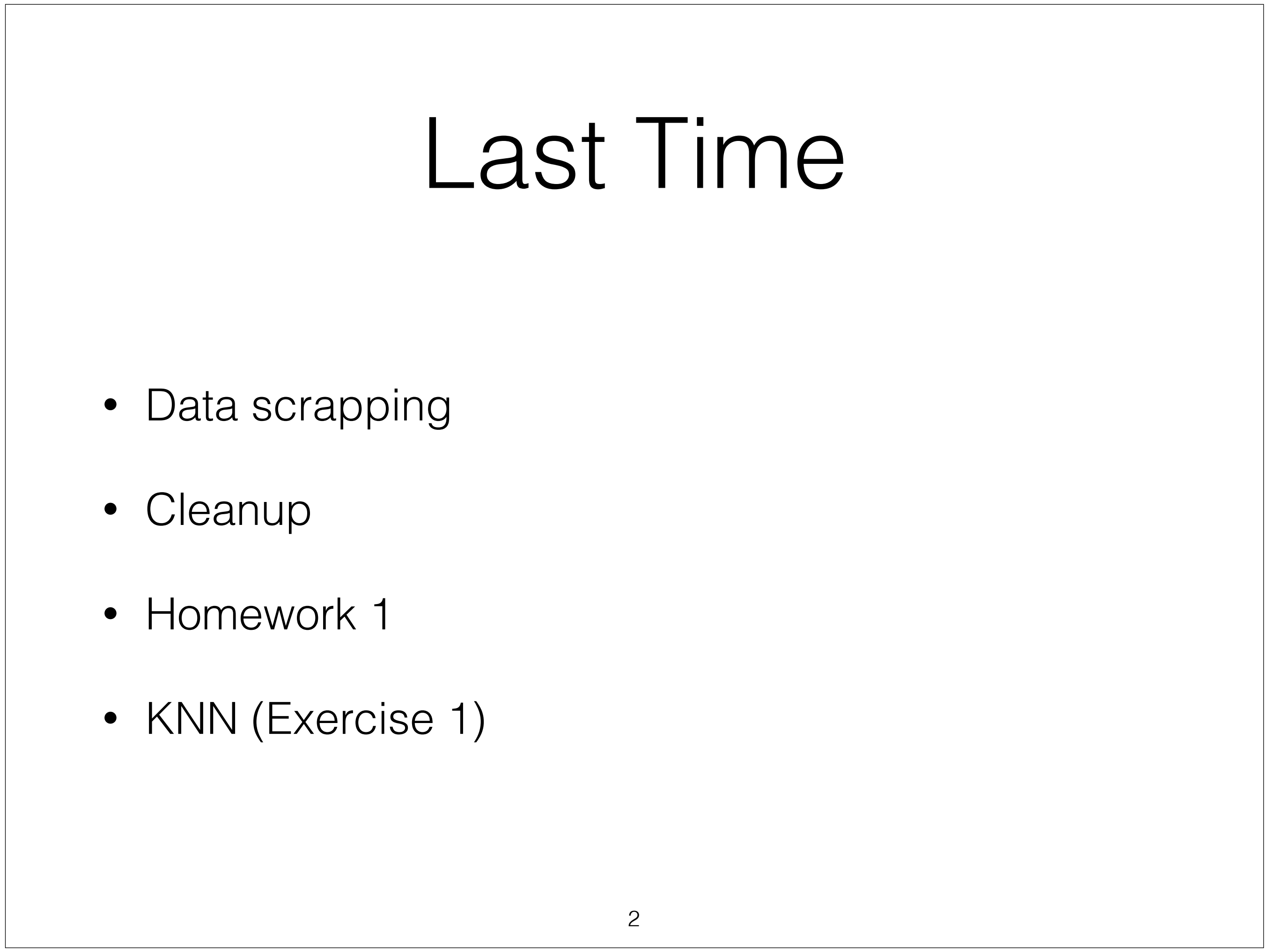
Freie Universität Berlin

Data Science Lecture 03

08.05.2015

Dr. Kashif Rasul
kashif @krasul #167

Welcome to Data Science Lecture 3.



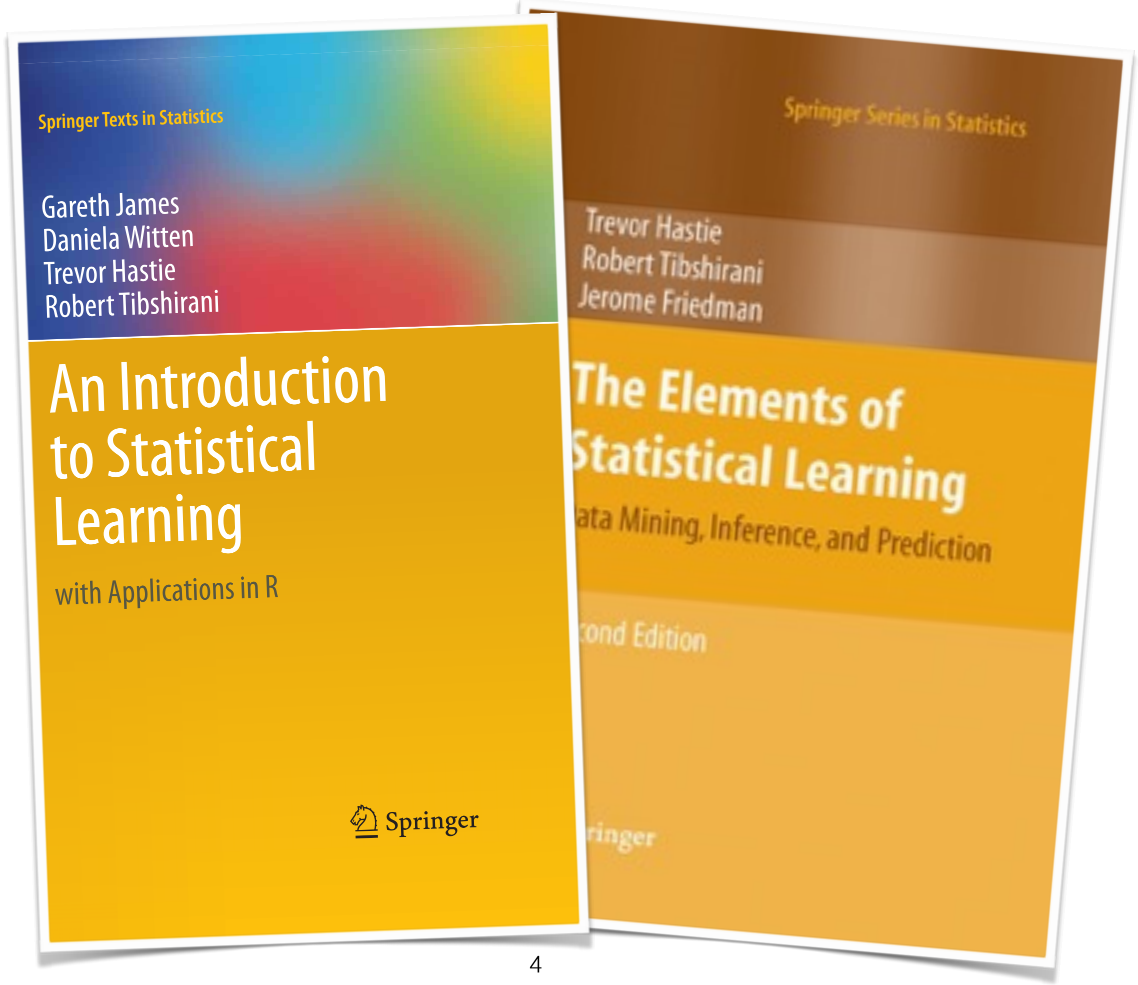
Last Time

- Data scrapping
- Cleanup
- Homework 1
- KNN (Exercise 1)

We hope you have made a start on Homework 1. Use the Github issues for asking questions and hopefully your colleagues can answer them too. As we mentioned you get credit for asking and answering questions which include Github issues.

Today

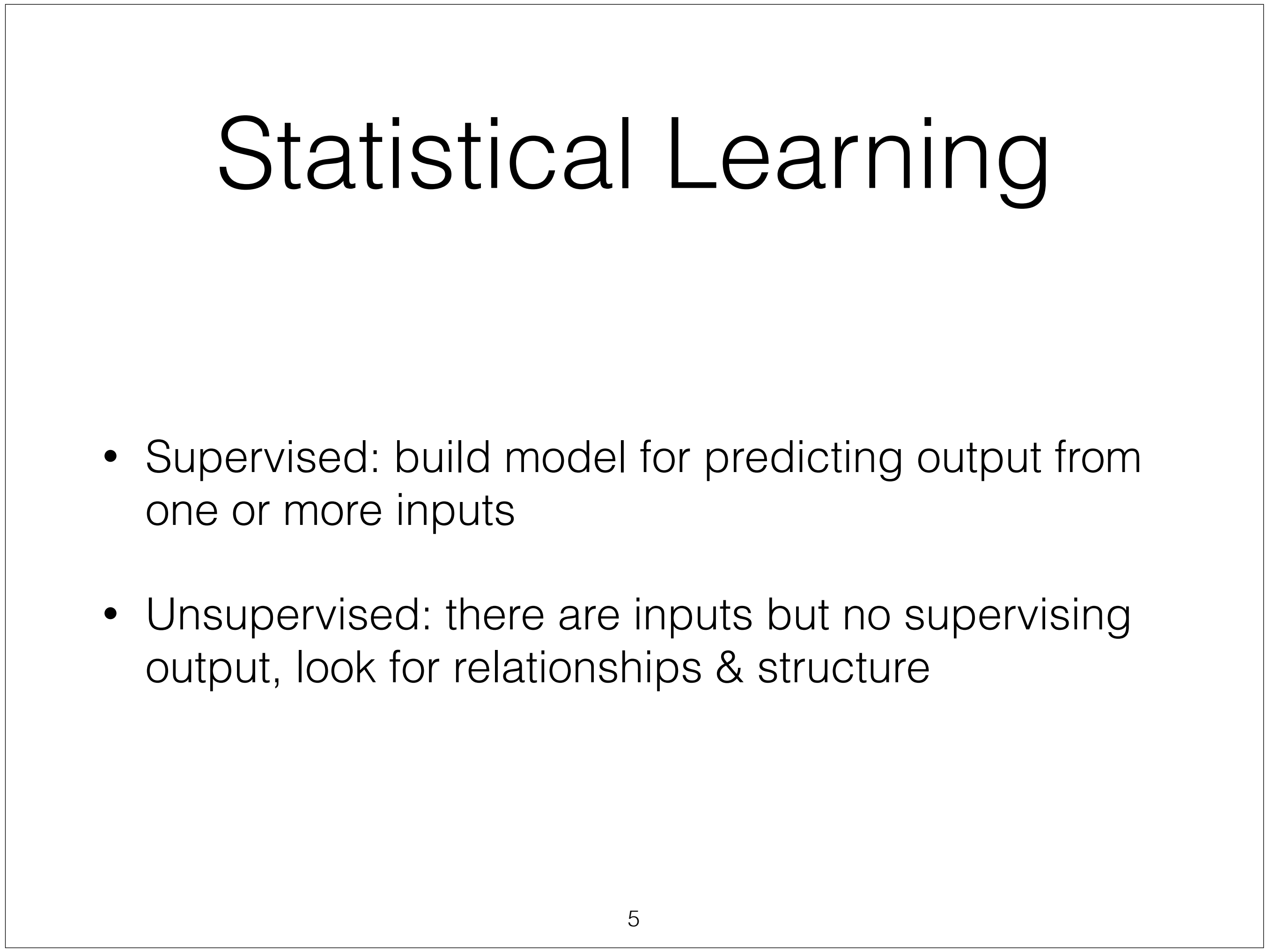
- Overview of Statistical Learning
- Bias-Variance Tradeoff
- Linear Regression


Springer Texts in Statistics
Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani
An Introduction to Statistical Learning
with Applications in R
Springer
Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
ata Mining, Inference, and Prediction
cond Edition
ringer

We will now use this book as our main reference. Download the book as pdf from: http://web.stanford.edu/~hastie/local.ftp/Springer/ISLR_print4.pdf

Note: Some of the figures in these slides are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

The second book is a more rigorous and advanced treatment of this first one and is available from: http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf



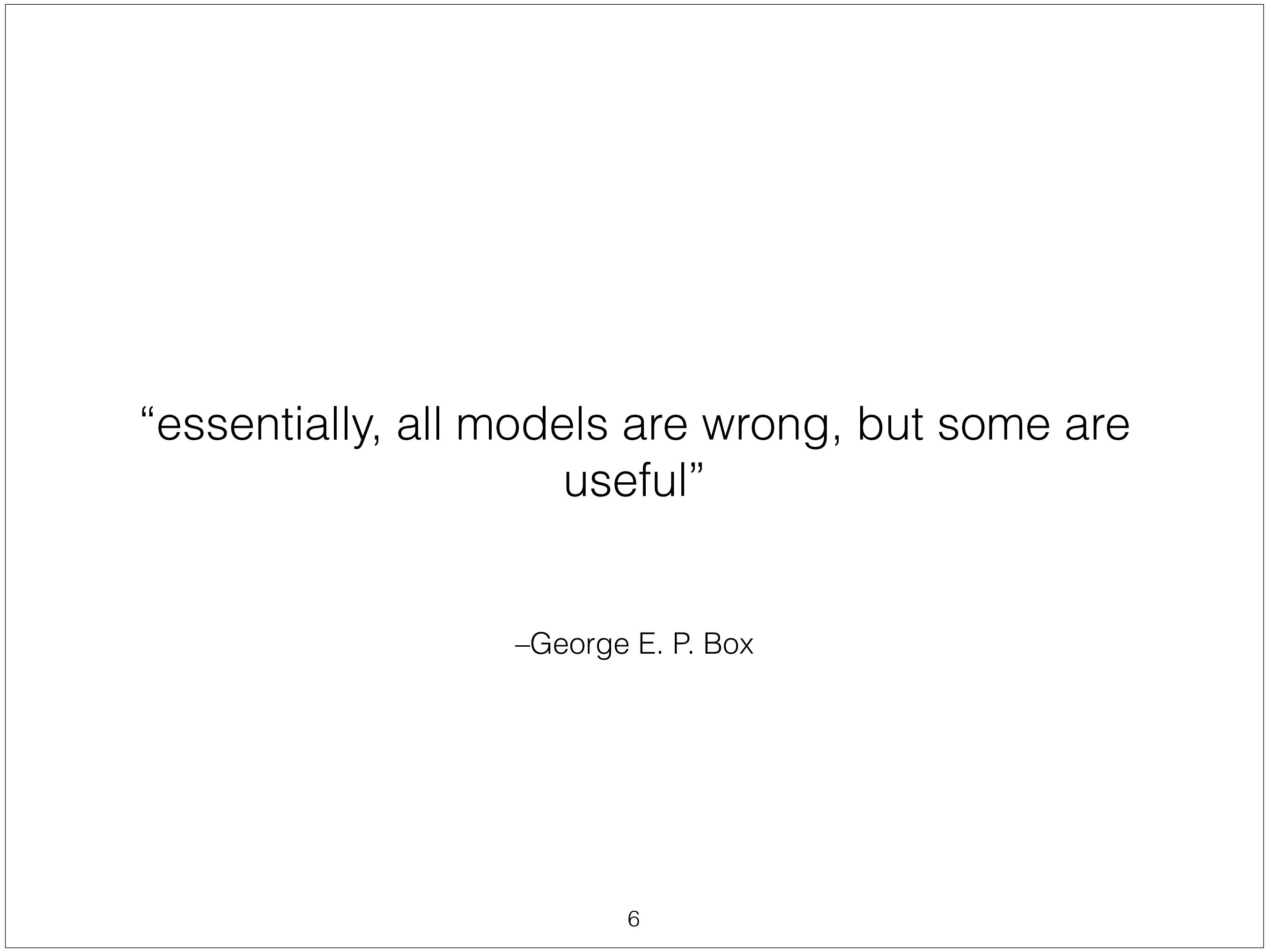
Statistical Learning

- Supervised: build model for predicting output from one or more inputs
- Unsupervised: there are inputs but no supervising output, look for relationships & structure

Statistical learning refers to a vast set of tools for understanding data. These tools can be classified as supervised or unsupervised.

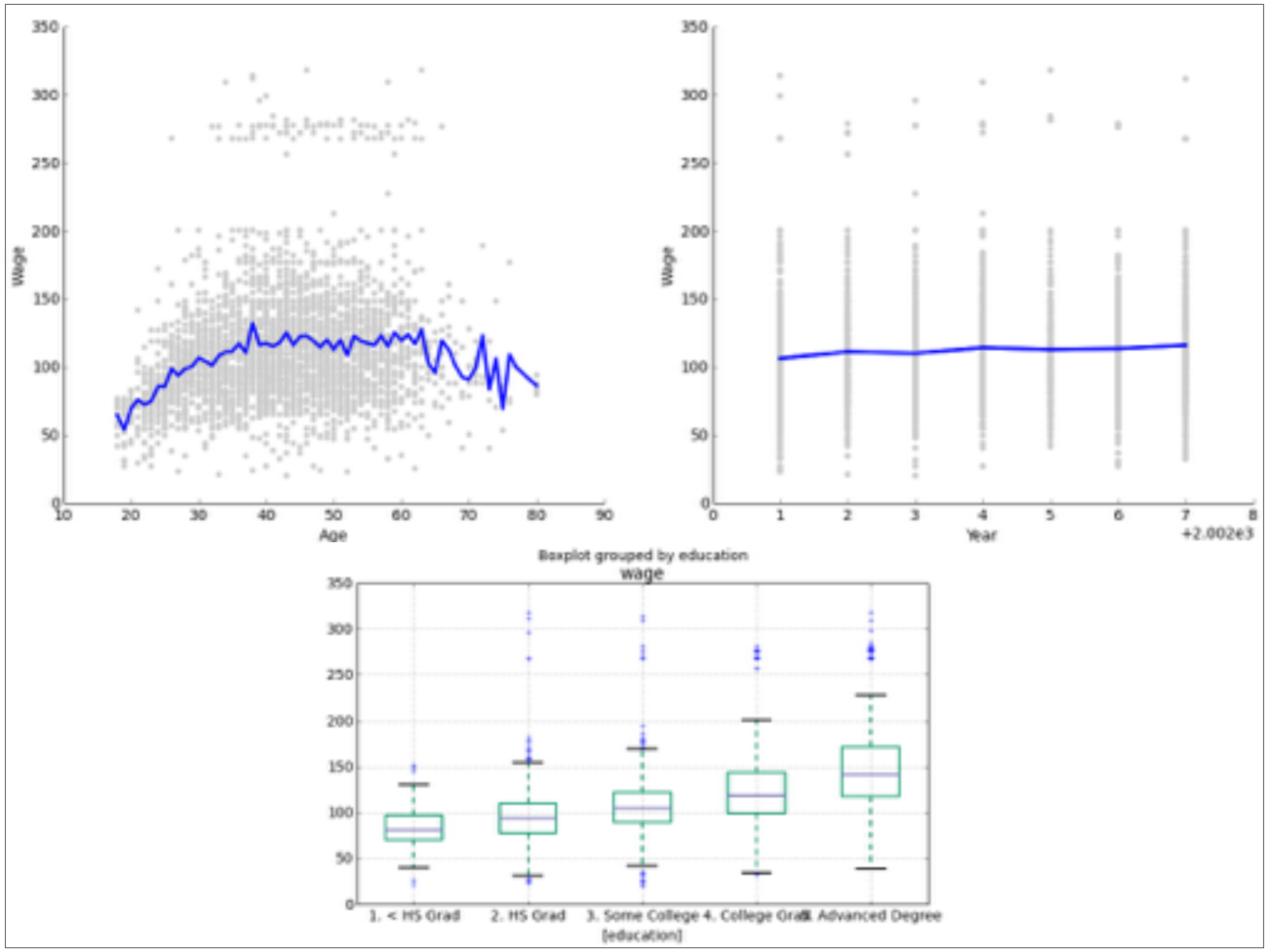
Supervised statistical learning involves building a statistical model for predicting, or estimating, an output based on one or more inputs. Problems of this nature occur in fields as diverse as business, medicine, astrophysics, and public policy.

Unsupervised statistical learning, there are inputs but no supervising output; nevertheless we can learn relationships and structure from such data.



> “essentially, all models are wrong, but some are useful”
>
> –George E. P. Box

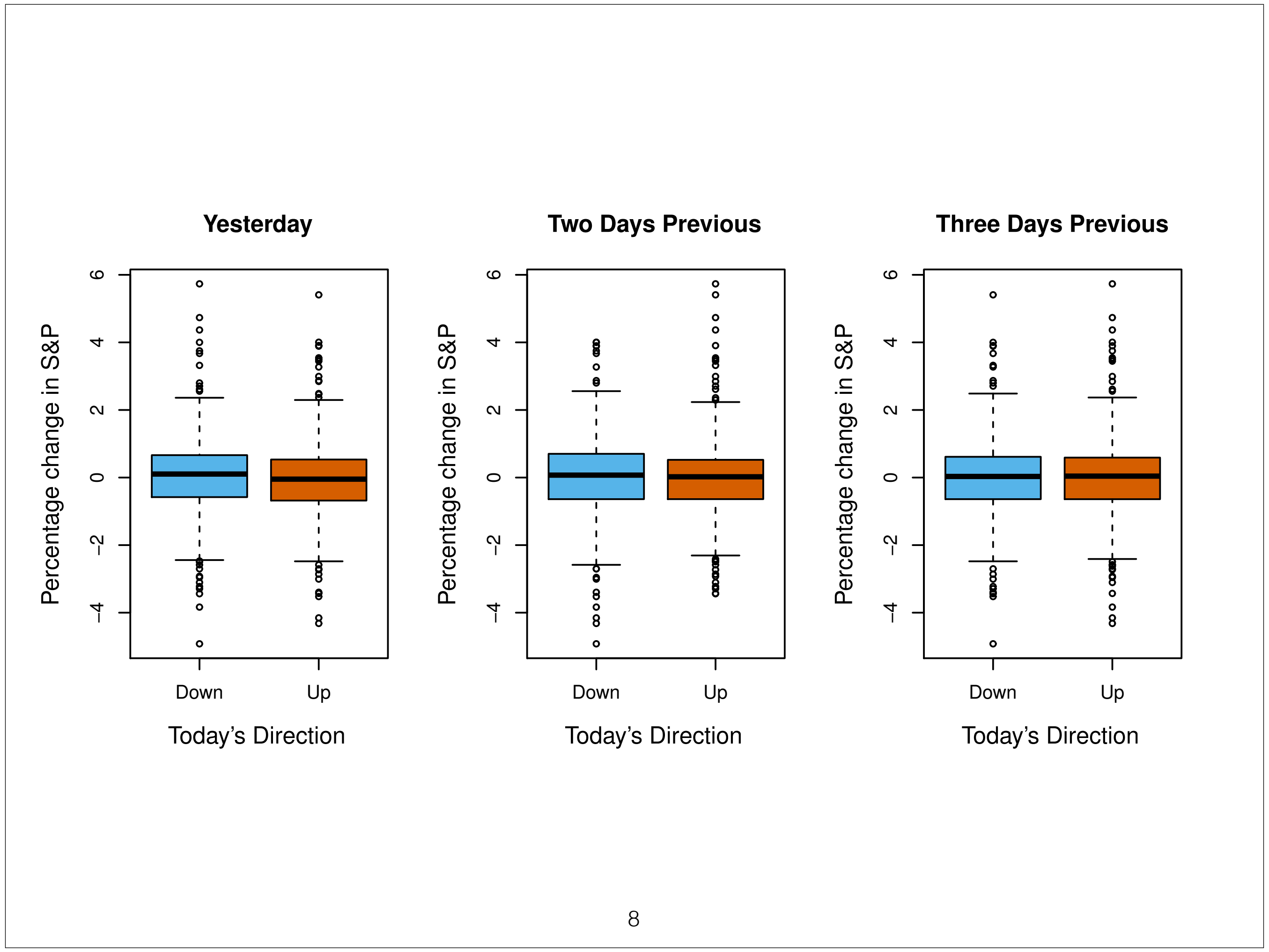
George Box a famous statistician in his book on response surface methodology wrote:



Let's look at the Wage data for a group of males from the Atlantic region of the United States. Consider wage versus age for each of the individuals in the data set. Given an employee's age, we can use say a curve around the means to predict his wage. However it is clear that there is a significant amount of variability associated with this average value, and so age alone is unlikely to provide an accurate prediction of a particular man's wage.

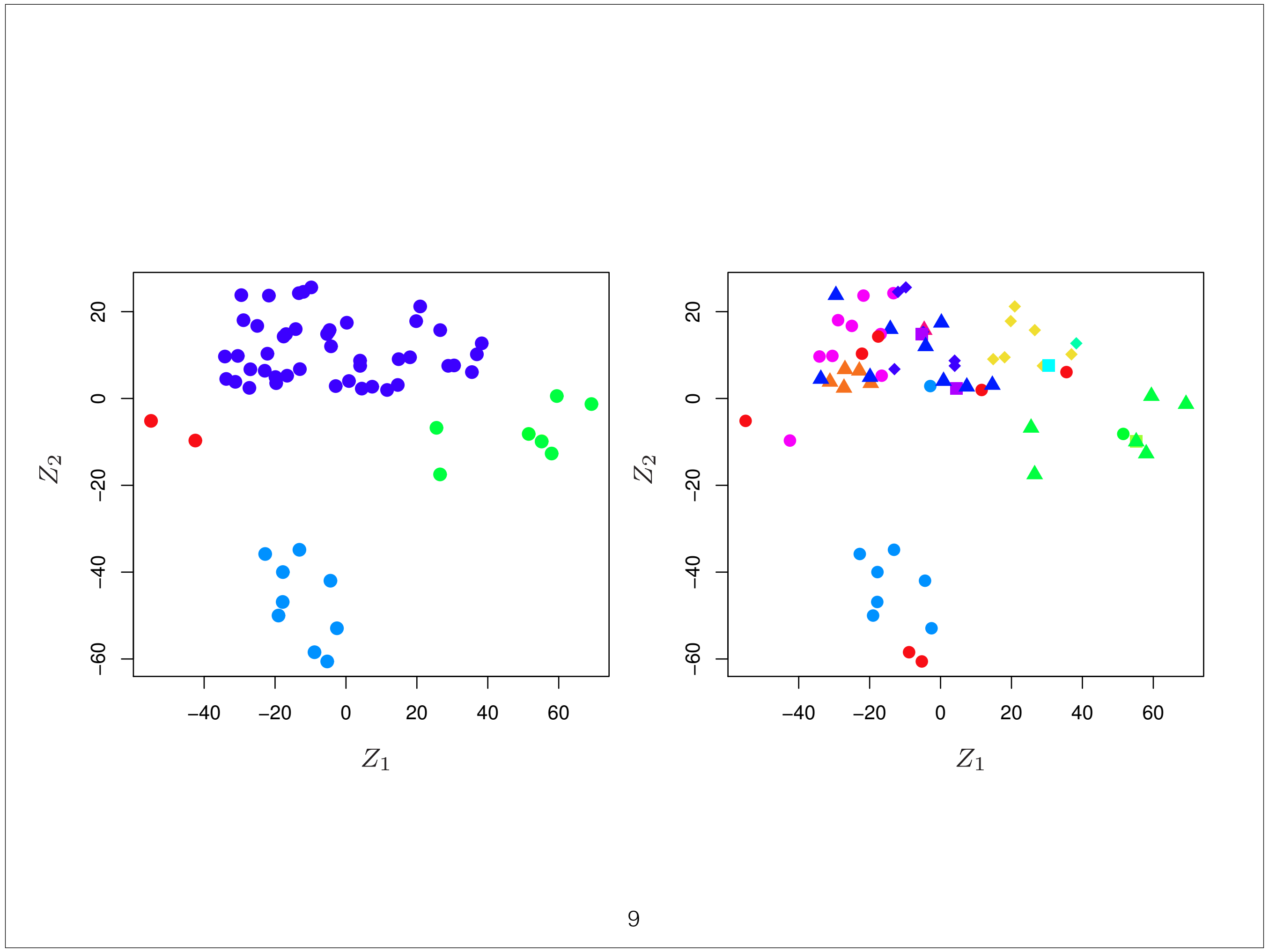
We also have information regarding each employee's education level and the year in which the wage was earned. Wages increase by approximately $10,000, in a roughly linear fashion. Wages are also typically greater for individuals with higher education levels.

Clearly, the most accurate prediction of a given man's wage will be obtained by combining his age, his education, and the year. This is referred to as a regression problem.



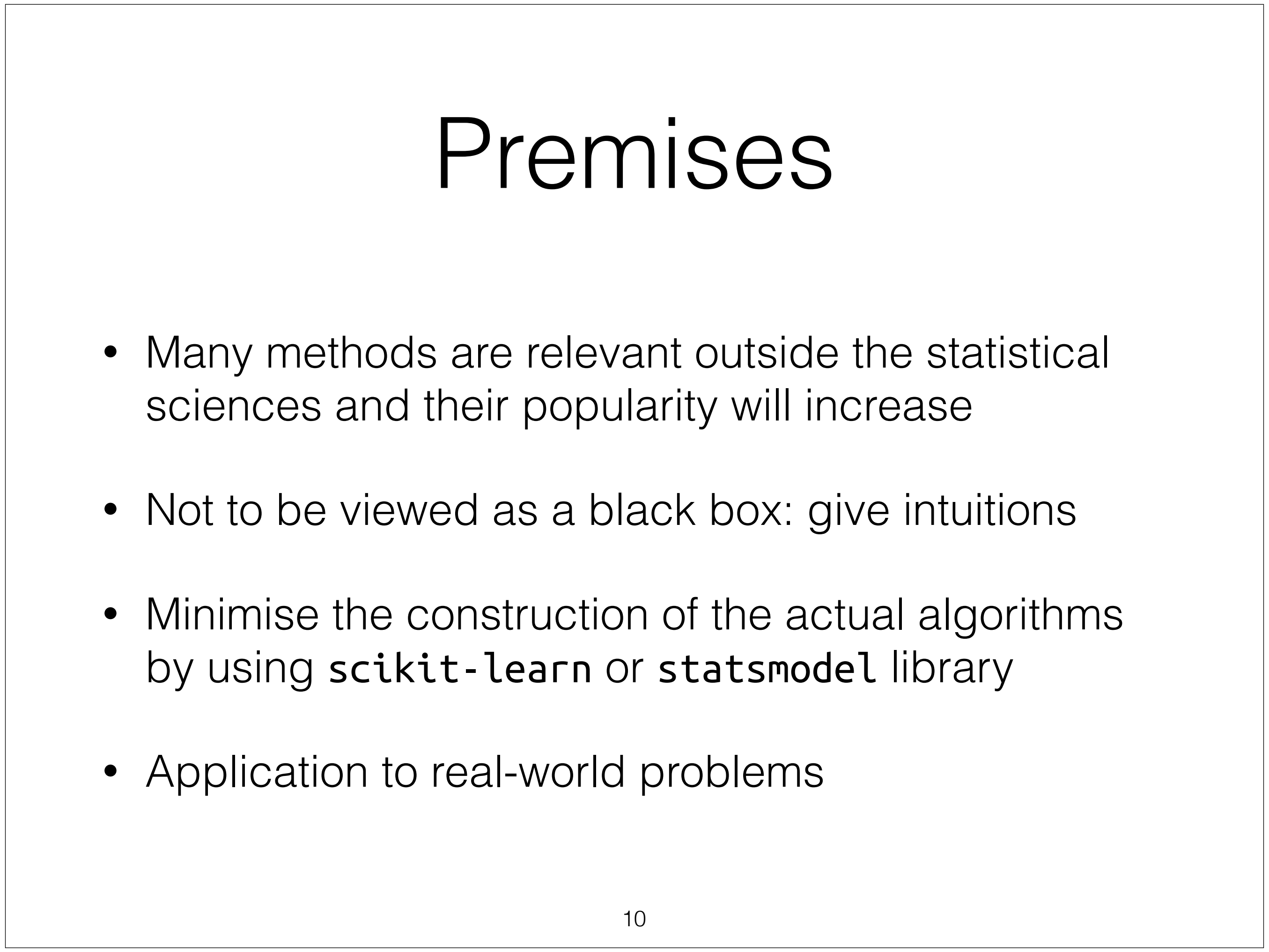
Wage data involves predicting a continuous value. However in certain cases we may instead wish to predict a non-numerical value, i.e. categorical output. In the SMarket data, the goal is to predict whether the index will increase or decrease on a given day using the past 5 days' percentage changes in the index.

Here we do not predict a value, but rather predict if something will fall into the 'up' bucket or 'down' bucket. This is known as a classification problem. Spam detection is another such example.



Here are two gene expression data in 2d from the NCI60 data set. Each point corresponds to one of the 64 cell lines. Three seems to be 4 groups of cell lines. The same data now represented with 14 different cancer types. Cell lines corresponding to the same cancer type tend to be nearby in the two-dimensional space.

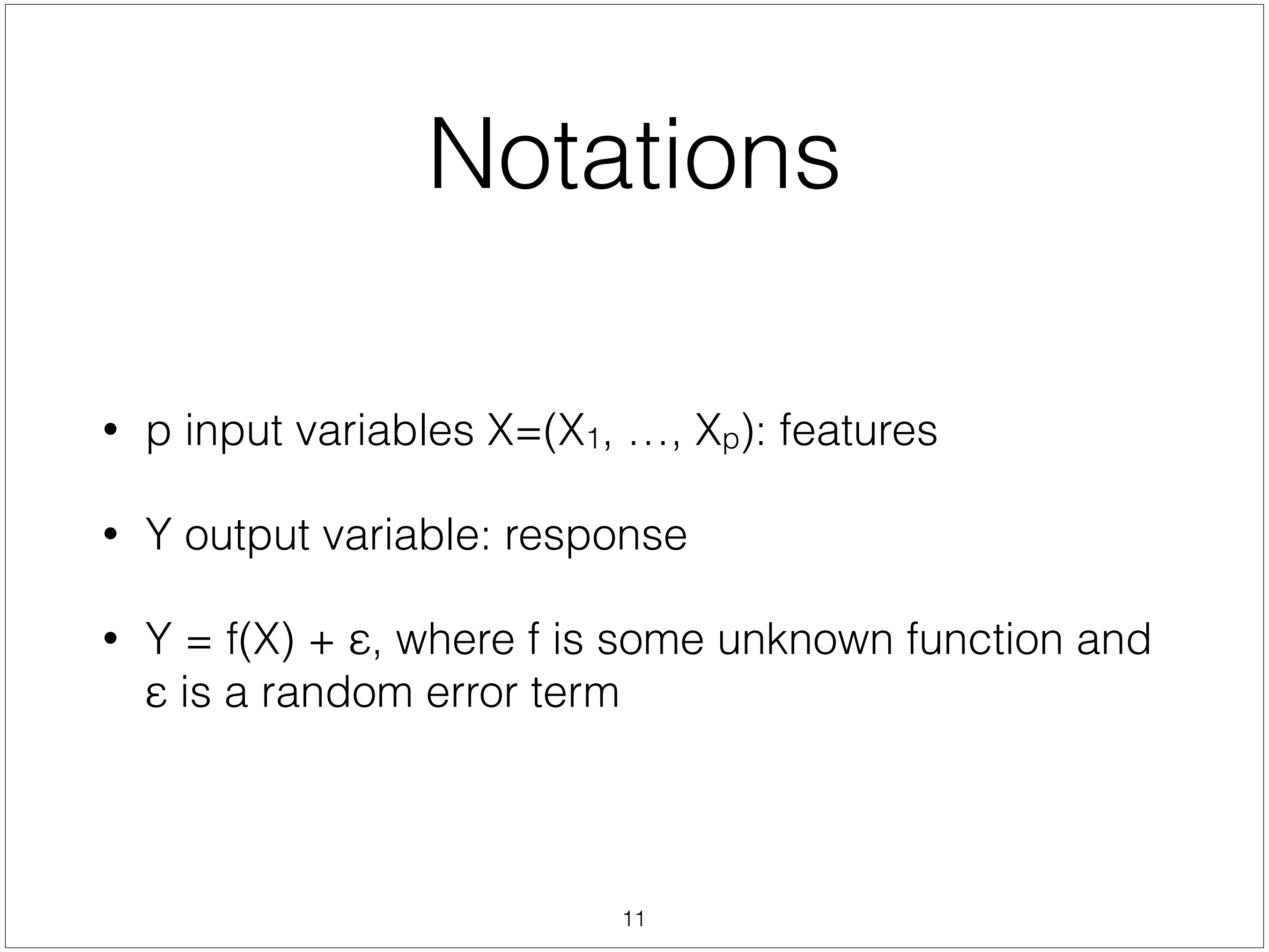
This is known as a clustering problem and we are not tying to predict an output.



Premises

- Many methods are relevant outside the statistical sciences and their popularity will increase
- Not to be viewed as a black box: give intuitions
- Minimise the construction of the actual algorithms by using `scikit-learn` or `statsmodel` library
- Application to real-world problems

Statistical learning is based on the following four premises:

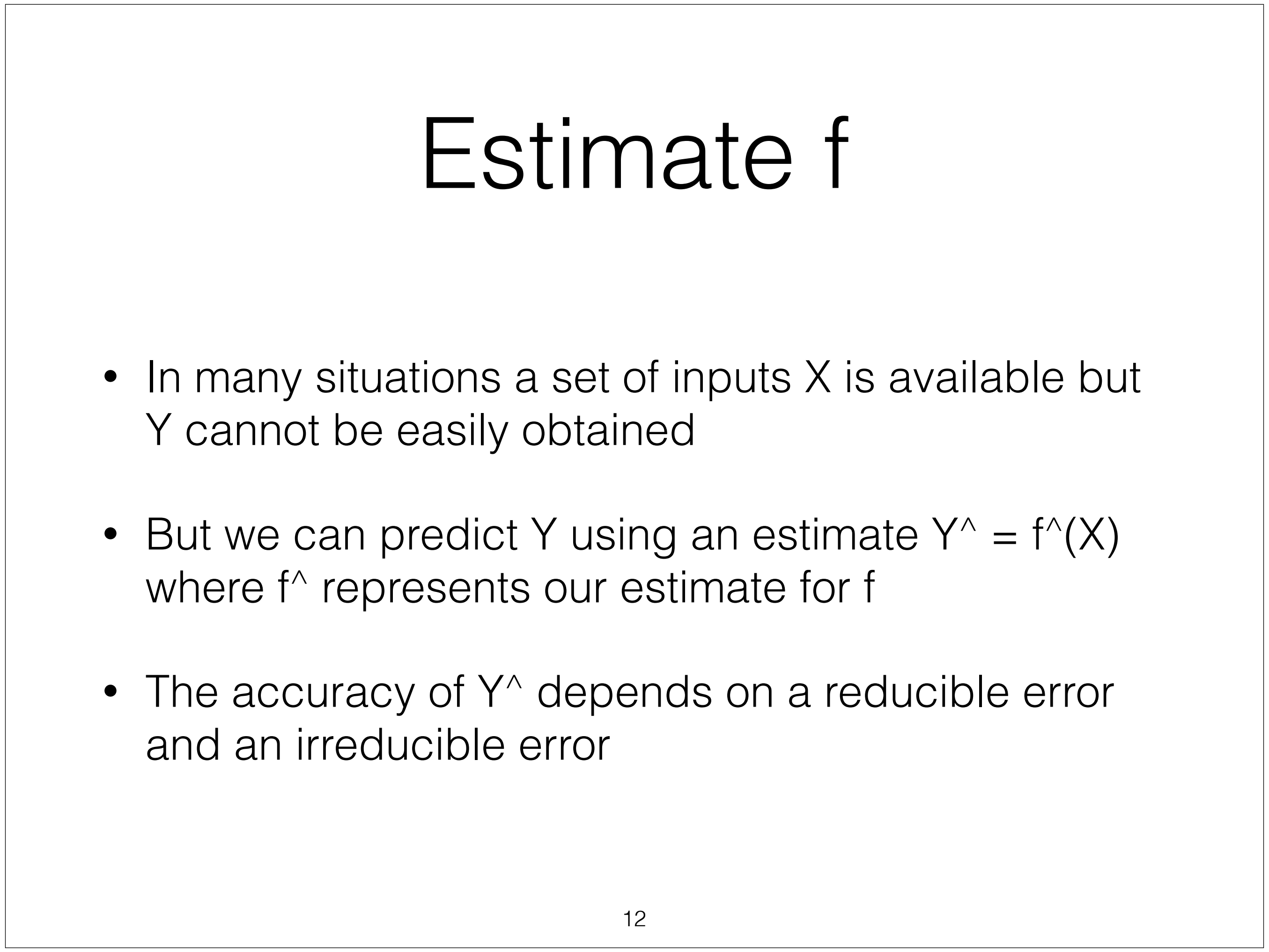


Notations

- p input variables $X=(X_1, \ldots, X_p)$: features
- Y output variable: response
- $Y = f(X) + \varepsilon$, where f is some unknown function and ε is a random error term

We assume that there is some relationship between Y and X represented by the general form.

In essence, statistical learning refers to a set of approaches for estimating f. There are two main reasons that we may wish to estimate f.



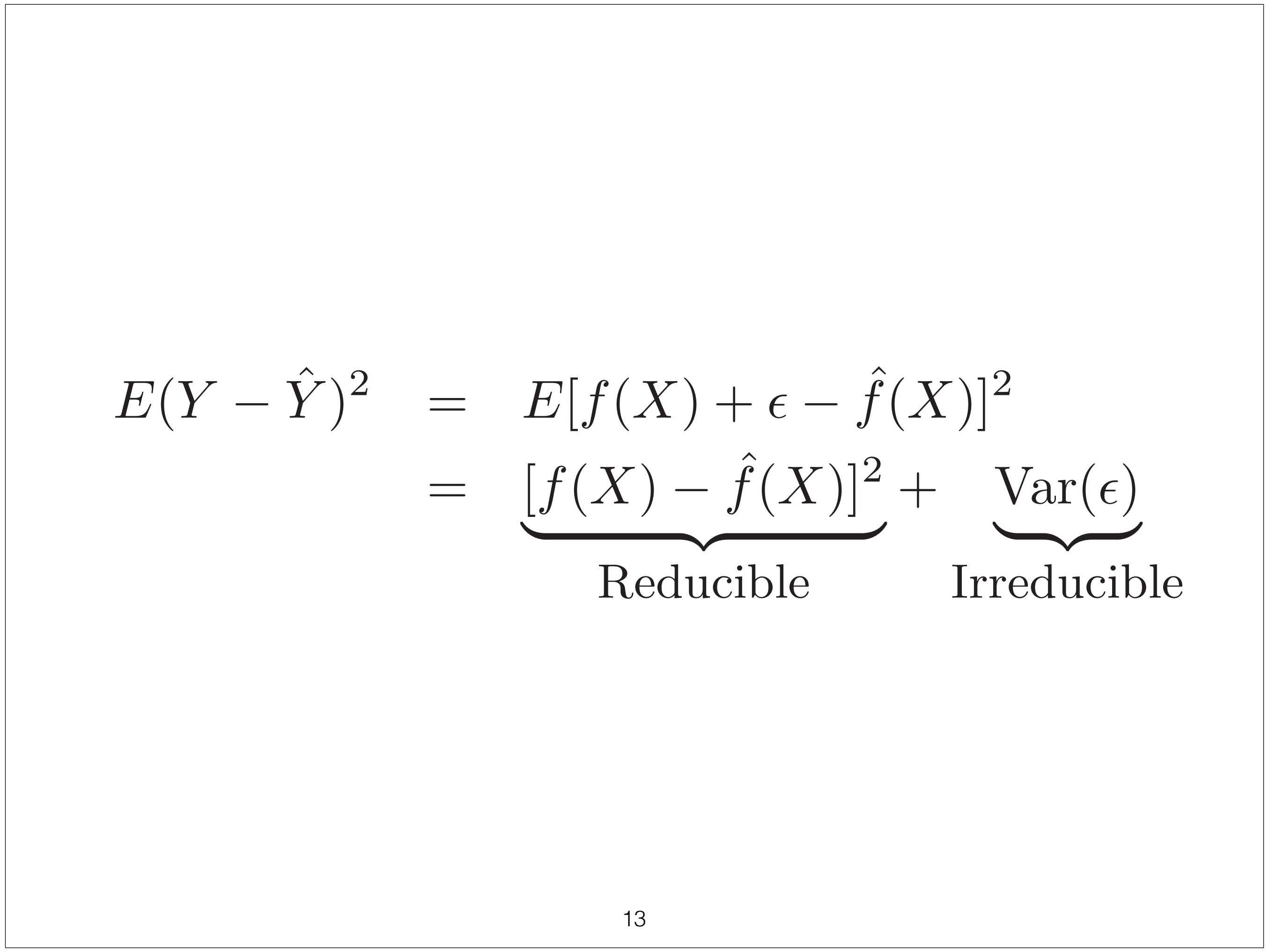
Estimate f

- In many situations a set of inputs X is available but Y cannot be easily obtained
- But we can predict Y using an estimate Y^ = f^(X) where f^ represents our estimate for f
- The accuracy of Y^ depends on a reducible error and an irreducible error

The first is prediction. In this setting we assume the error term averages to zero.

We also don’t care about the exact form of $\hat{f}$ provided that it yields accurate predictions for Y.

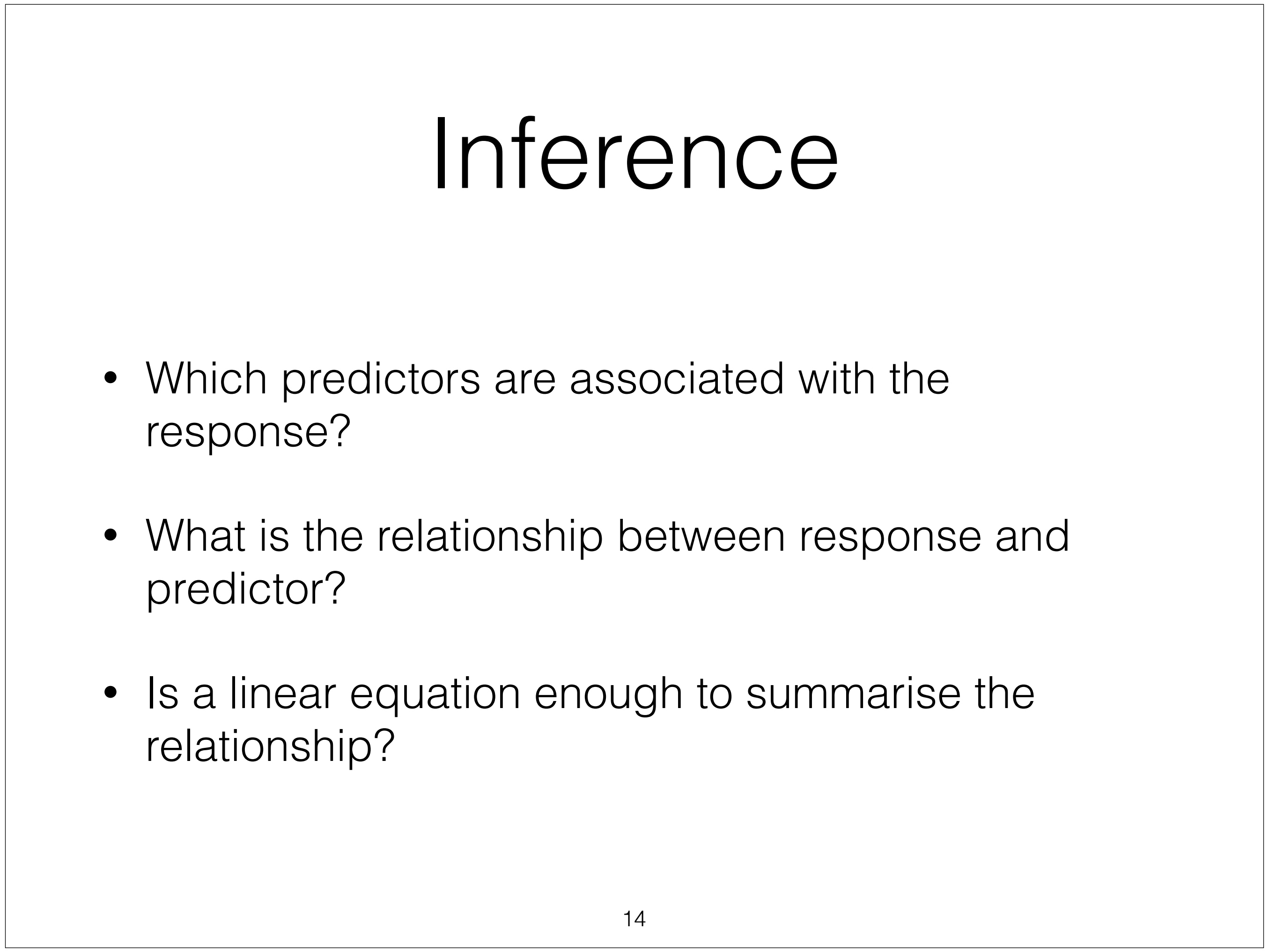
The error is reducible because we can potentially improve the accuracy of our model. But our model will still have some error in it. This is because Y is also a function of ε which by definition cannot be predicted using X. Therefore variability associated with ε also affects the accuracy of our predictions. This is the source of the irreducible error.



$$\begin{aligned} E(Y-\hat{Y})^2 &= E[f(X)+\epsilon-\hat{f}(X)]^2 \\ &= \underbrace{[f(X)-\hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}} \end{aligned}$$

Consider a given estimate f^ and a set of predictors X, which yields a prediction Y^. Assuming for a moment that both f^ and X are fixed then we have the above where $E(Y-\hat{Y})^2$ is the average or expected value of the squared difference between the actual and predicted value of Y and Var(ε) is the variance associated with the error term ε.

The main aim of statistical learning is to minimise the reducible error.



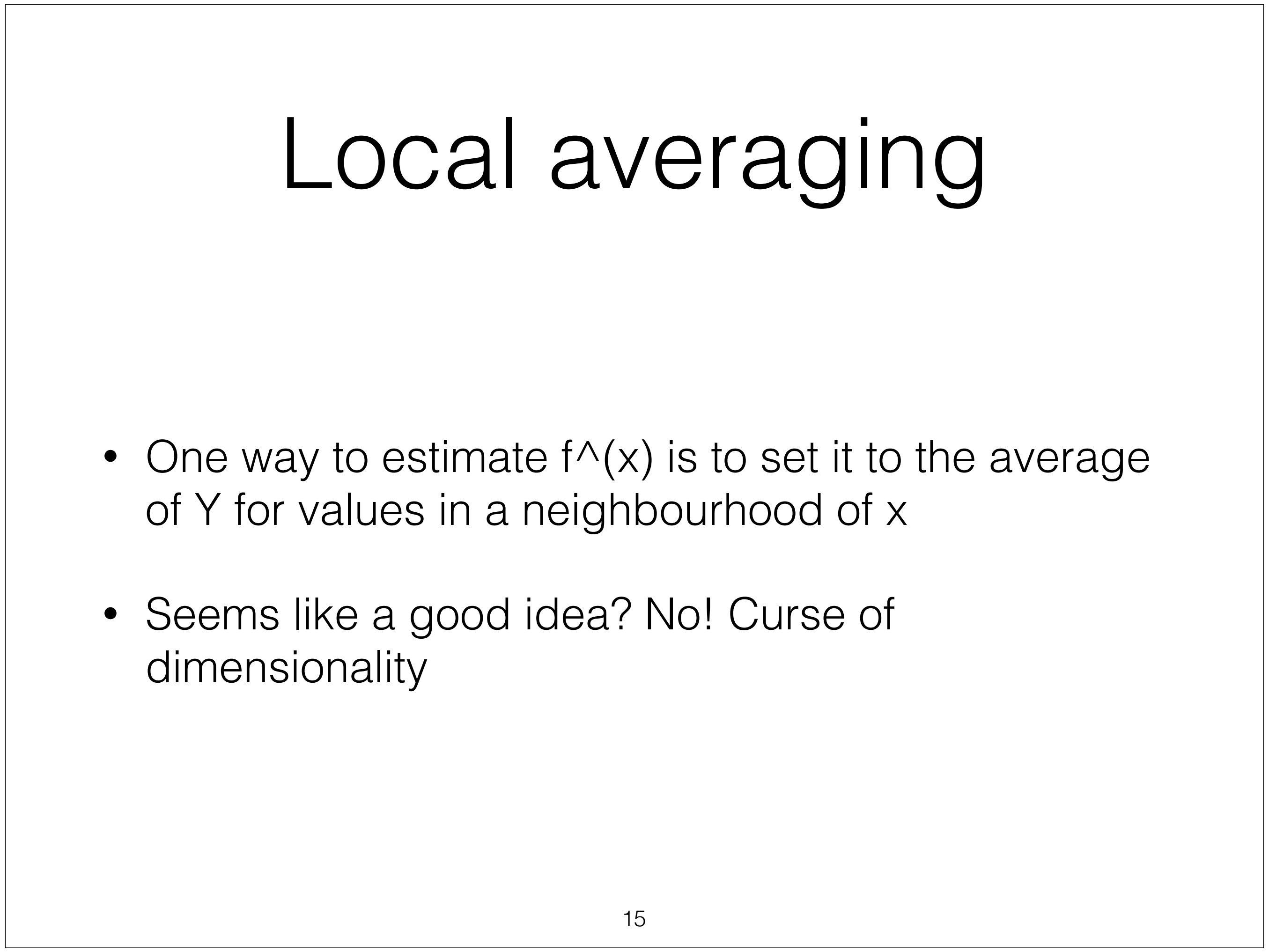
Inference

- Which predictors are associated with the response?
- What is the relationship between response and predictor?
- Is a linear equation enough to summarise the relationship?

The second reason is inference. We are often interested in the way Y is affected as X changes. So we wish to estimate f but our goal is not to make an accurate prediction, instead we want to understand the relationship between X and Y. Now the predictor cannot be treated as a black box since we wish to know its exact form.

In this setting one is interested in:

In general linear models allow for relatively simple and interpretable inference but may not yield accurate predictions as some other approaches.

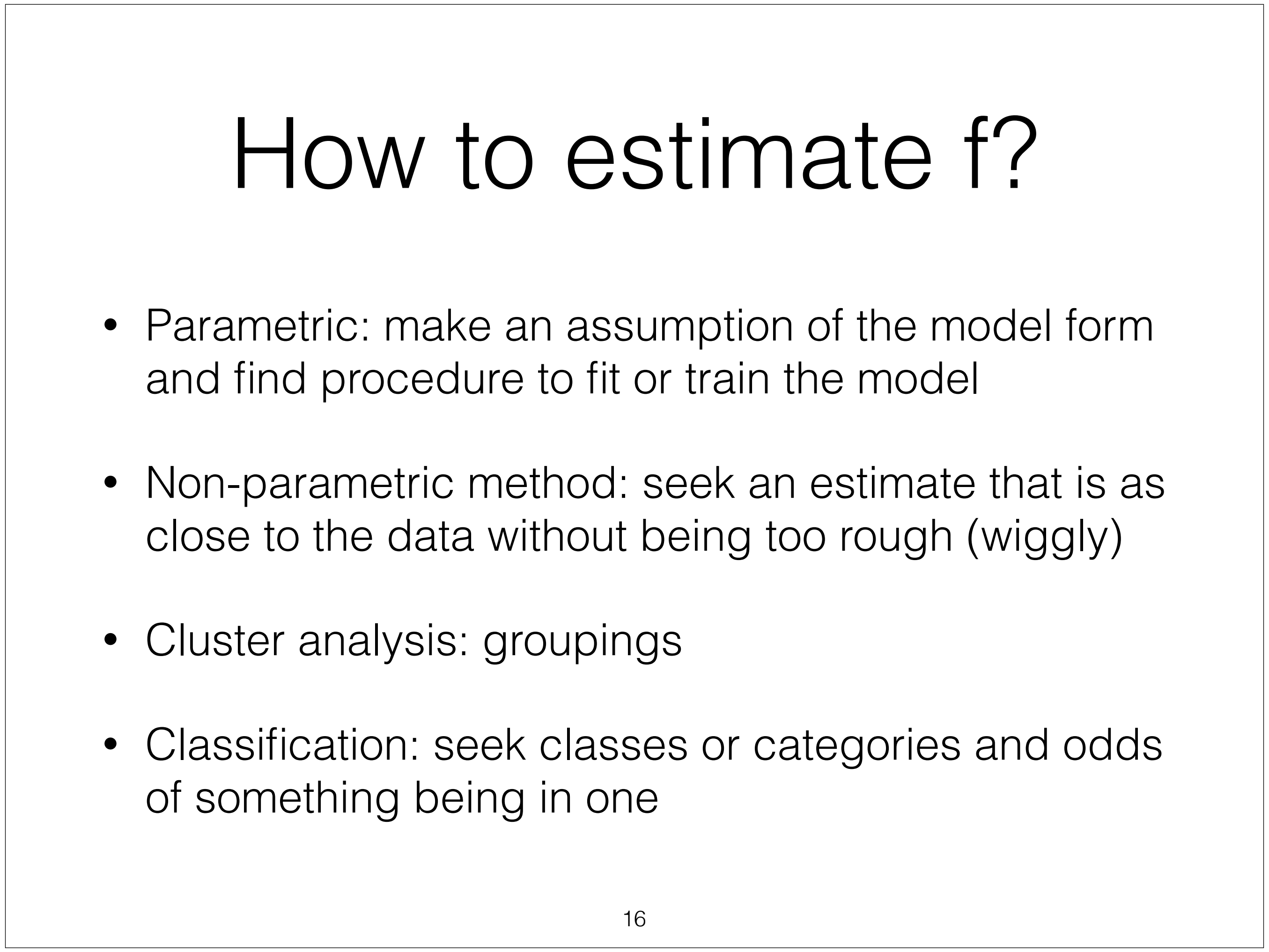


Local averaging

- One way to estimate f^(x) is to set it to the average of Y for values in a neighbourhood of x
- Seems like a good idea? No! Curse of dimensionality

Nearest neighbour averaging doesn't really work so well. The problem has to do with the curse of dimensionality. It works well for small p<=4 and large N where we have a lot of data points to take averages from. This is one class of techniques called smoothers and we will discuss more sophisticated methods later.

When p is large, neighbours tend to be far away… e.g. a 10% neighbourhood in high dimensions is no longer local and we end up averaging everything…



How to estimate f?

- Parametric: make an assumption of the model form and find procedure to fit or train the model
- Non-parametric method: seek an estimate that is as close to the data without being too rough (wiggly)
- Cluster analysis: groupings
- Classification: seek classes or categories and odds of something being in one

How to choose between these methods? Well when inference is the goal, the linear model may be a good choice since it will be quite easy to understand the relationship between Y and X. In general the more flexible the method, the less interpretable it becomes.

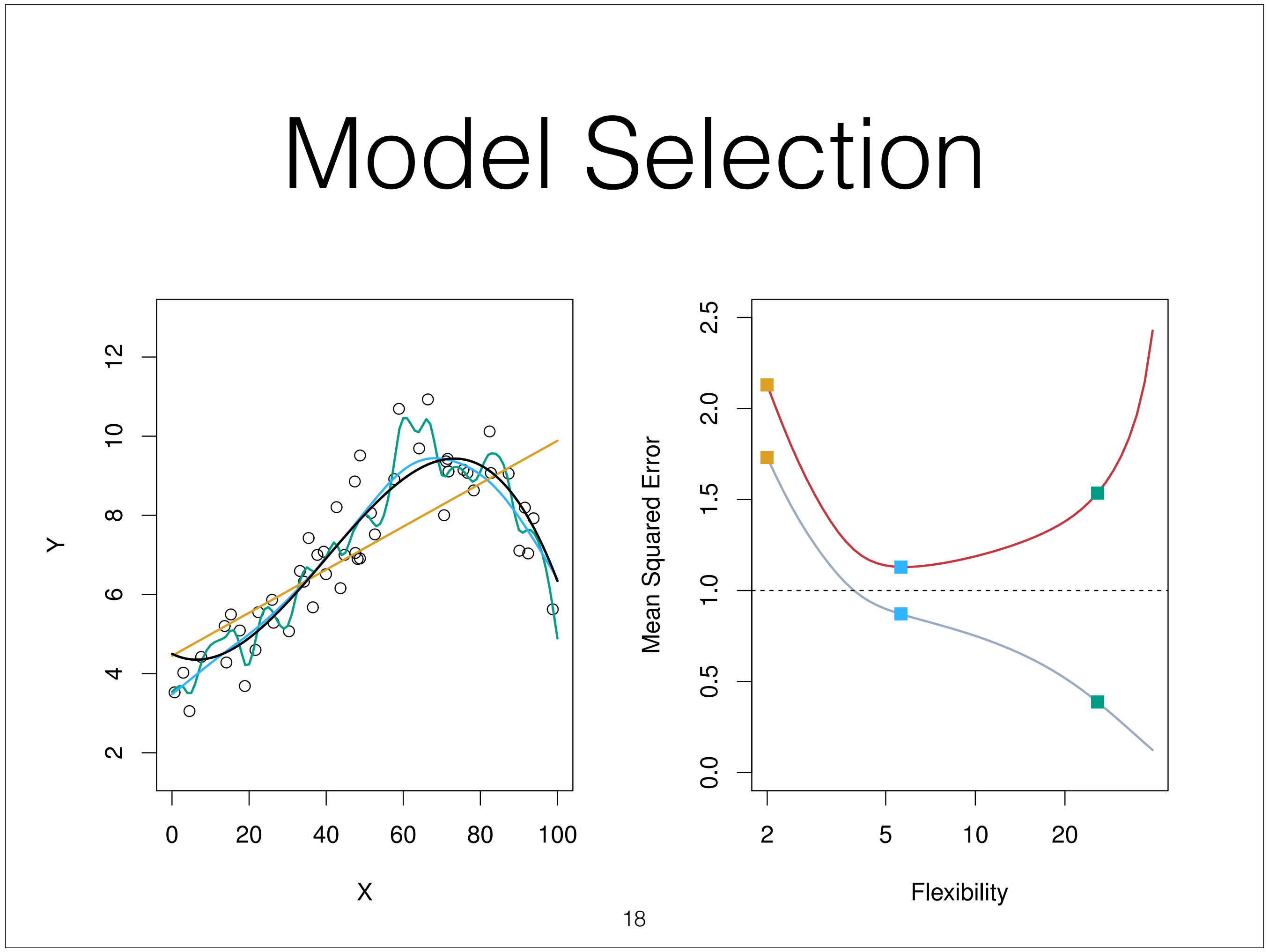
Note these methods are for the supervised setting and unsupervised learning is a bit more challenging. One tool is cluster analysis.

Finally variables can be quantitative or qualitative . For classification problems the odds of some qualitative variable belonging to a class can be generalised by using binary variables and regression etc.

Trade-offs

- Prediction accuracy vs. Interpretability
- Good fit verses under-fit or over-fit
- Parsimony vs. black-box

Parsimony means having a model that is simpler and that can be transmitted with a small number of parameters and explained in a simple fashion.



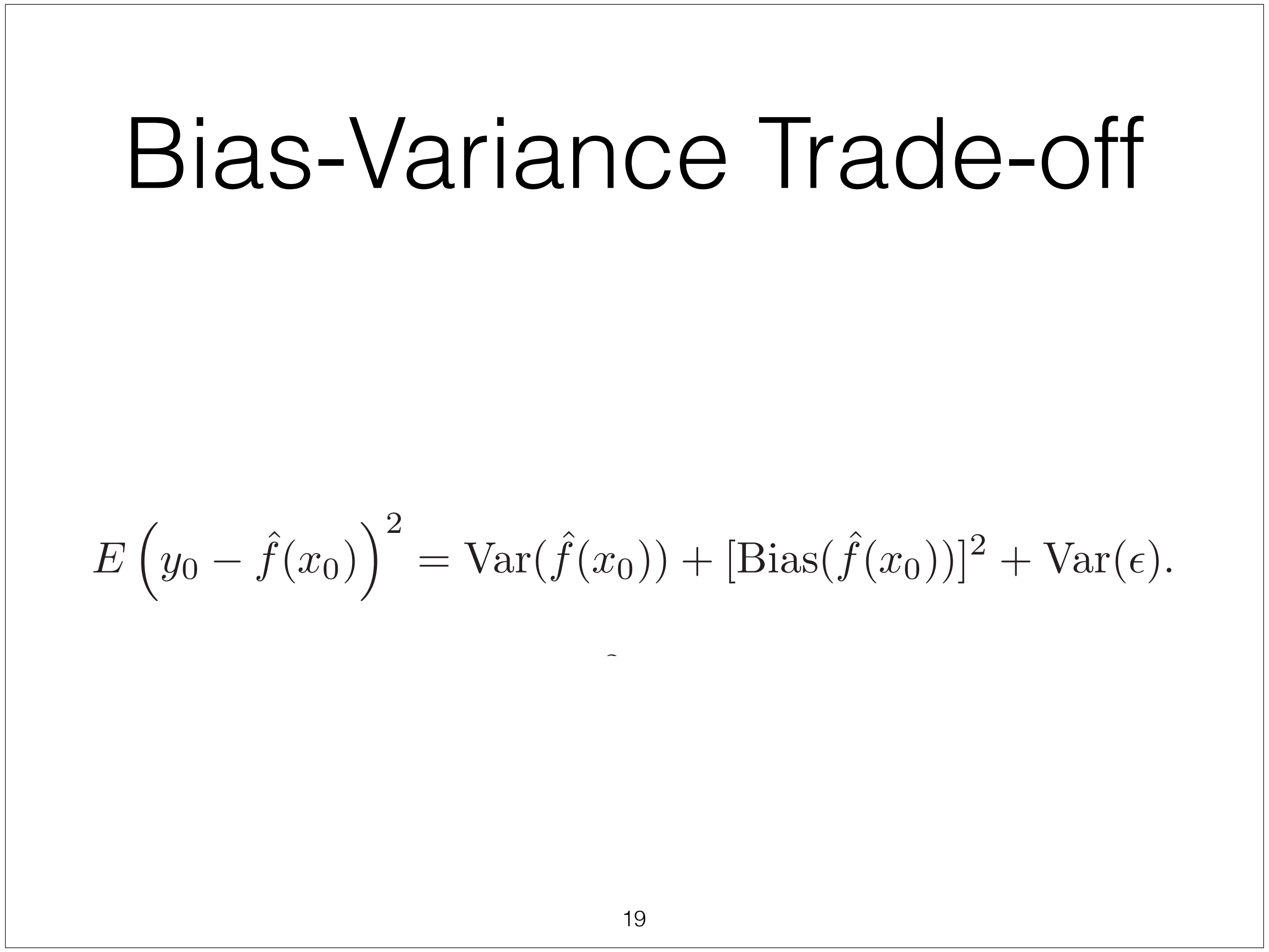
Model Selection

Suppose we fit a model f (x) to some data and wish to see how well it performs. We could compute the average square prediction error or MSE over this data. But this may be bias towards more overfit models.

Instead if we could compute the average over fresh test data and compared these different models. We see that, test MSE decreases initially but as the flexibility increases it starts to increase as well. Since this is generated data, the dash-line is the Var(ε).

This is a fundamental property of statistical learning that holds regardless of the particular data set at hand and regardless of the statistical method being used. When we overfit the test MSE will be very large.

Later we will discuss ways to find this minimum point for selecting the best model.



Bias-Variance Trade-off

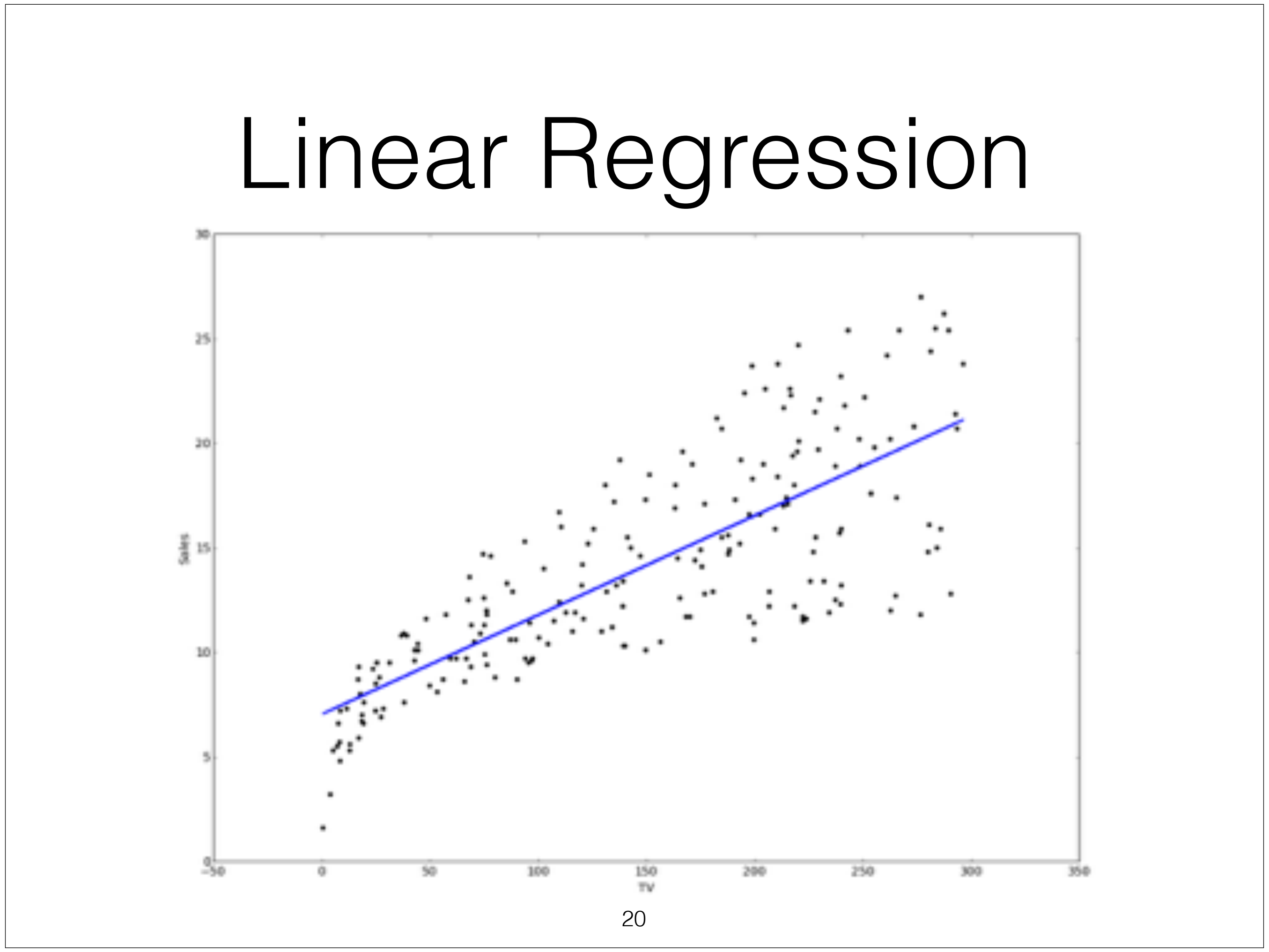
$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

The first fundamental concept of statistical learning is to get an intuition on the U shaped curve of the test MSE in the previous slide. This has to do with the fact that the MSE can be decomposed into three fundamental quantities:

Here the expected test MSE refers to the average test MSE that we would obtain if we repeatedly estimated f using a large number of training sets and and tested each at x_0. What this tells us is that in order to reduce the test error, we need a method with low variance and bias squared, both of which are positive.

Intuitively, variance is the amount $\hat{f}$ would change if we estimated it using different training data. So more flexible methods have higher variance.

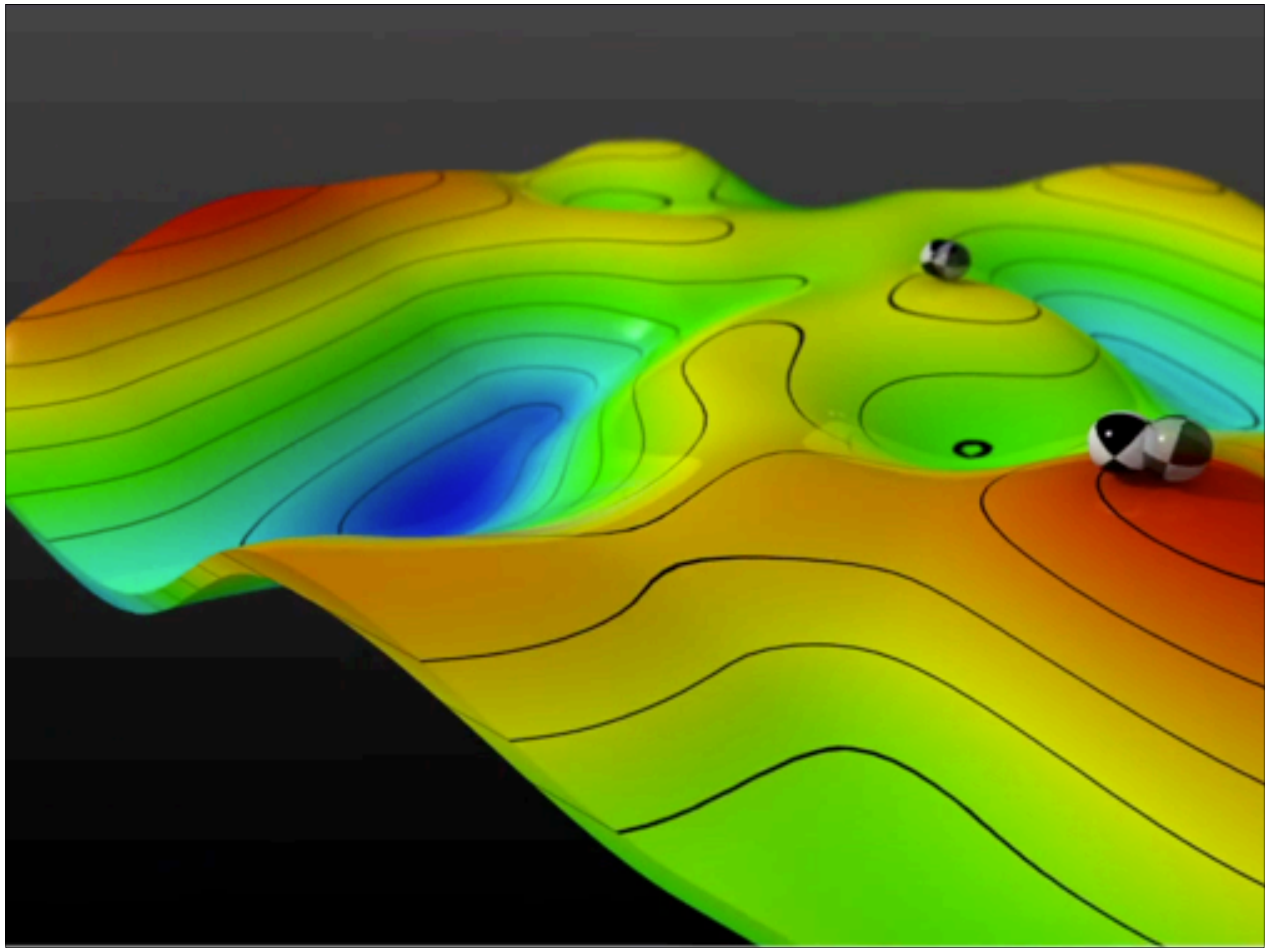
The bias refers to the error introduced by approximating real-life problems, which are complicated, by a simple model. E.g. linear regression assumes a linear relationship between Y and X. Real-life is not like that. So more flexible models have a lower bias.



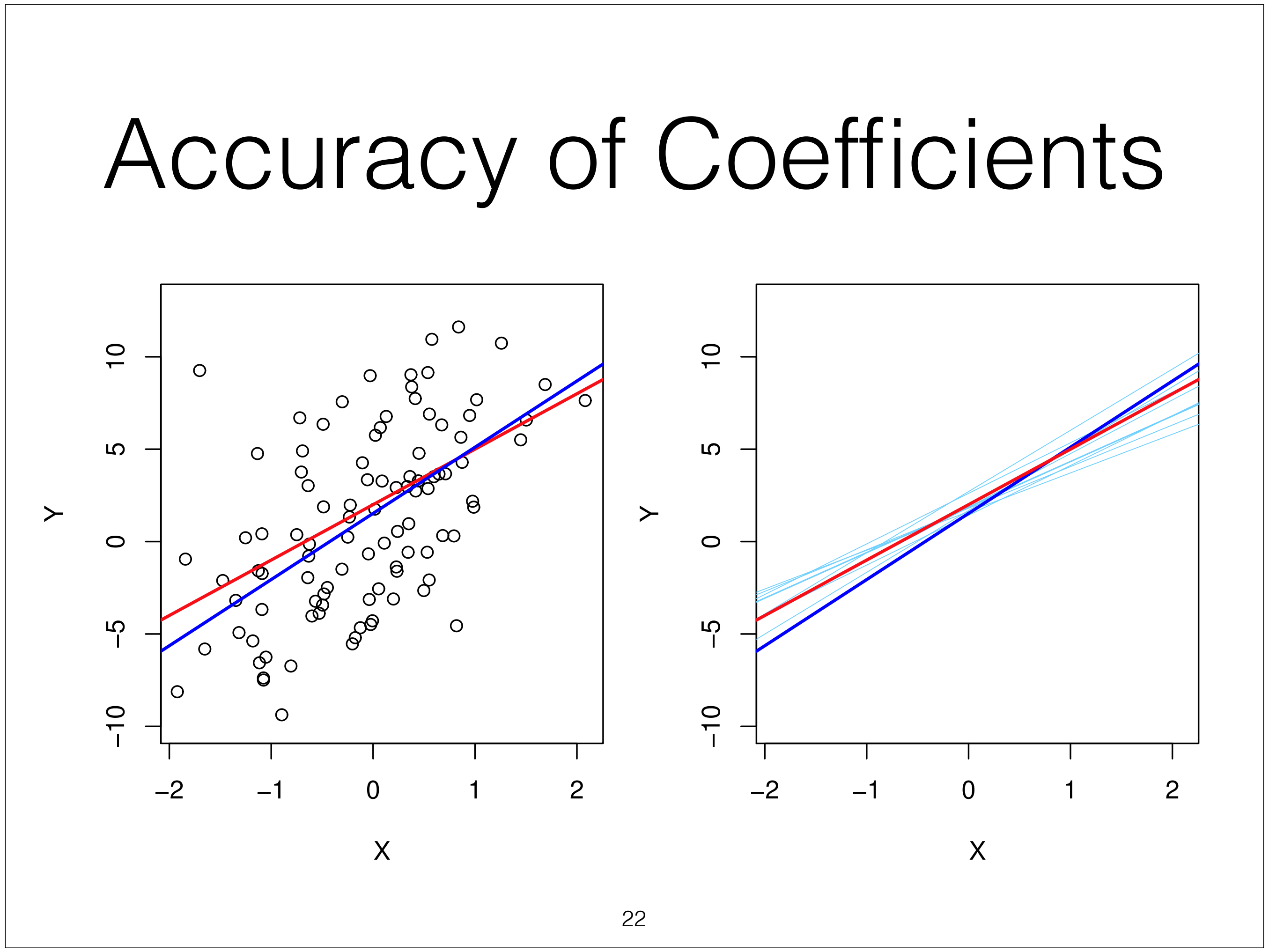
Linear Regression

So let’s discuss linear regression, a very simple approach for supervised learning to predicting a quantitative response. For the advertising data, we can regress sales≈β_0 +β_1 ×TV for some unknown β_0 and β_1.

The idea is to use least squares in order to reduce the Residual sum of squares.



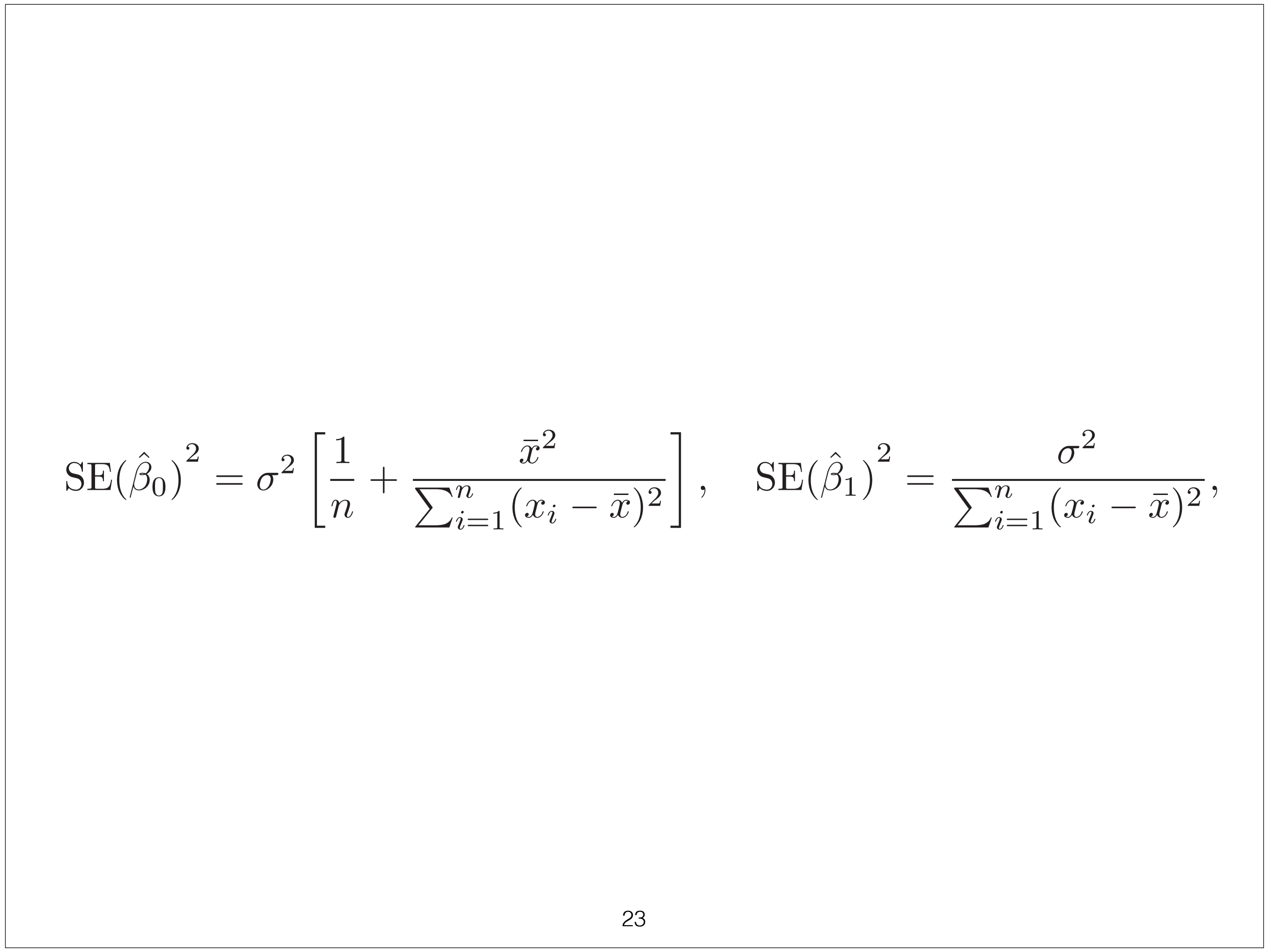
One algorithm to find the parameters that reduce the residuals is to use gradient descent. We start with an initial guess for the parameter and repeatedly update the parameter in the direction of the steepest descent of the cost function, with some fixed size step till some convergence.



Accuracy of Coefficients

To discuss the accuracy of the coefficients lets simulate some data from a known function shown in red in the left diagram. The least squares estimate is in black. Then on the right we show the least square in dark blue and least squares for subsets of the data. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.

The sample mean $\hat{\mu}$ is unbiased meaning on average we expect it to be equal to the true population mean. The same holds for the coefficients, that on average the regression line is close to the true model.

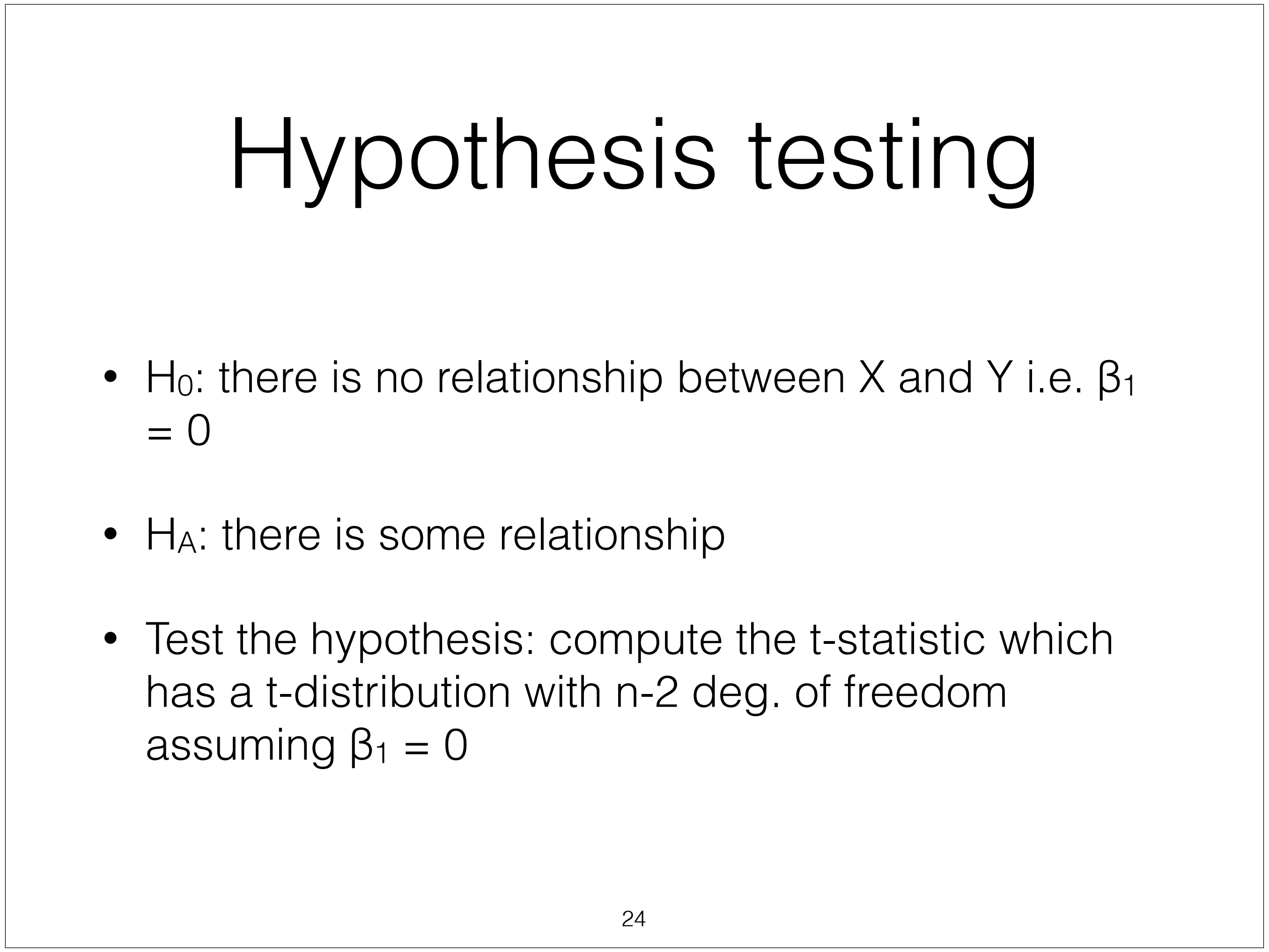


$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

But how far off is that single estimate μ ? We can compute the standard error of μ , SE(μ). We have Var(μ) = SE(μ) = σ /n where σ is the std. dev. of each y_i. Intuitively the SE tells us the average amount that this estimate will differ from its actual value and it scales with 1/n, the more observations we have the smaller this gets.

Similarly we have the standard error of the coefficients where σ^2 = Var(ε) and we assume ε_i is uncorrelated with common variance σ^2. This is not true but the formula is a good approximation. Notice SE(β_1) is smaller when the x_i are more spread out. Also SE(β_0) = SE($\hat{\mu}$) when $\bar{x}$ = 0.

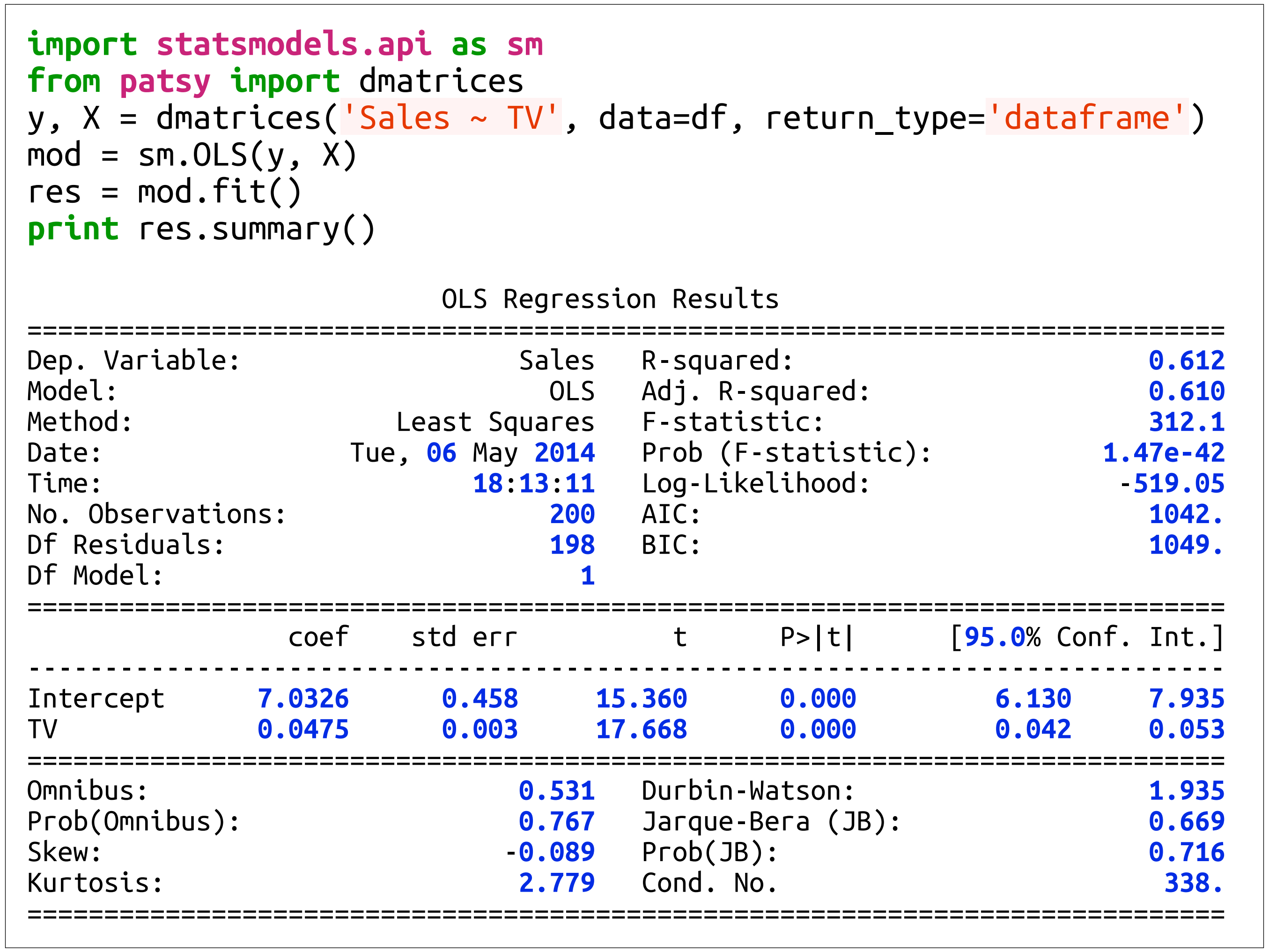
Since σ^2 is unknown we can estimate it with the residual std. error RSE = sqrt(RSS/(n-2)). Given this we can compute confidence intervals, e.g. he 95% confidence interval for β_0 is [6.130, 7.935] and the 95% confidence interval for β_1 is [0.042, 0.053]. So we can conclude that in the absence of any advertising, sales will, on average, fall somewhere between 6,130 and 7,940 units.



Hypothesis testing

- H_0: there is no relationship between X and Y i.e. $\beta_1 = 0$
- H_A: there is some relationship
- Test the hypothesis: compute the t-statistic which has a t-distribution with n-2 deg. of freedom assuming $\beta_1 = 0$

We can also compute the probability of observing any value equal to |t| or larger which is called the p-value. Intuitively a small p-value means that it is unlikely to observe such a substantial association between Y and X due to chance, in the absence of any real association between them. So a small p-value means that there is an association between the predictor and response, and we reject the null hypothesis.

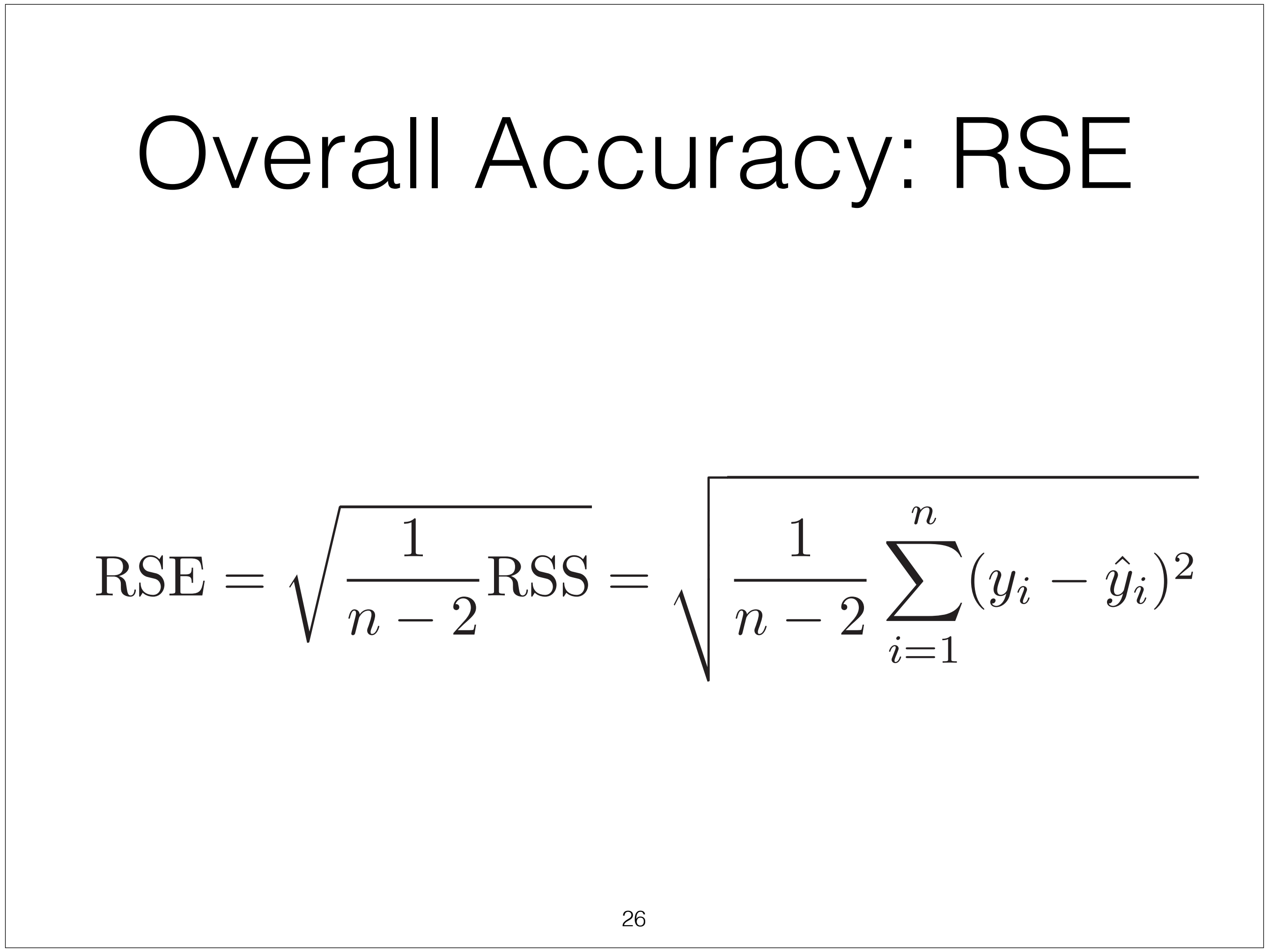


```python
import statsmodels.api as sm
from patsy import dmatrices
y, X = dmatrices('Sales ~ TV', data=df, return_type='dataframe')
mod = sm.OLS(y, X)
res = mod.fit()
print res.summary()
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.612
Model:                            OLS   Adj. R-squared:                  0.610
Method:                 Least Squares   F-statistic:                     312.1
Date:                Tue, 06 May 2014   Prob (F-statistic):           1.47e-42
Time:                        18:13:11   Log-Likelihood:                -519.05
No. Observations:                 200   AIC:                             1042.
Df Residuals:                     198   BIC:                             1049.
Df Model:                           1
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      7.0326      0.458     15.360      0.000         6.130     7.935
TV             0.0475      0.003     17.668      0.000         0.042     0.053
==============================================================================
Omnibus:                        0.531   Durbin-Watson:                   1.935
Prob(Omnibus):                  0.767   Jarque-Bera (JB):                0.669
Skew:                          -0.089   Prob(JB):                        0.716
Kurtosis:                       2.779   Cond. No.                         338.
==============================================================================
```

So for the advertising data and using just the TV here are the results. So we have t statistics = coef/(std err). We are not too interested in the Intercept numbers, but rather the t values for TV. The t-statistic is huge. In order to have a p value of 0.05 we need a t-statistic of 2. So 17.668 means it's very significant. Another way to say it is that it is very unlikely that TV has no effect on Sales.

The confidence interval and hypothesis testing are equivalent. If we reject the null hypothesis, then the confidence interval will not contain zero and conversely if we cannot reject the null hypothesis, then the conf. interval will contain zero. The confidence interval is also telling you how big the effect is.

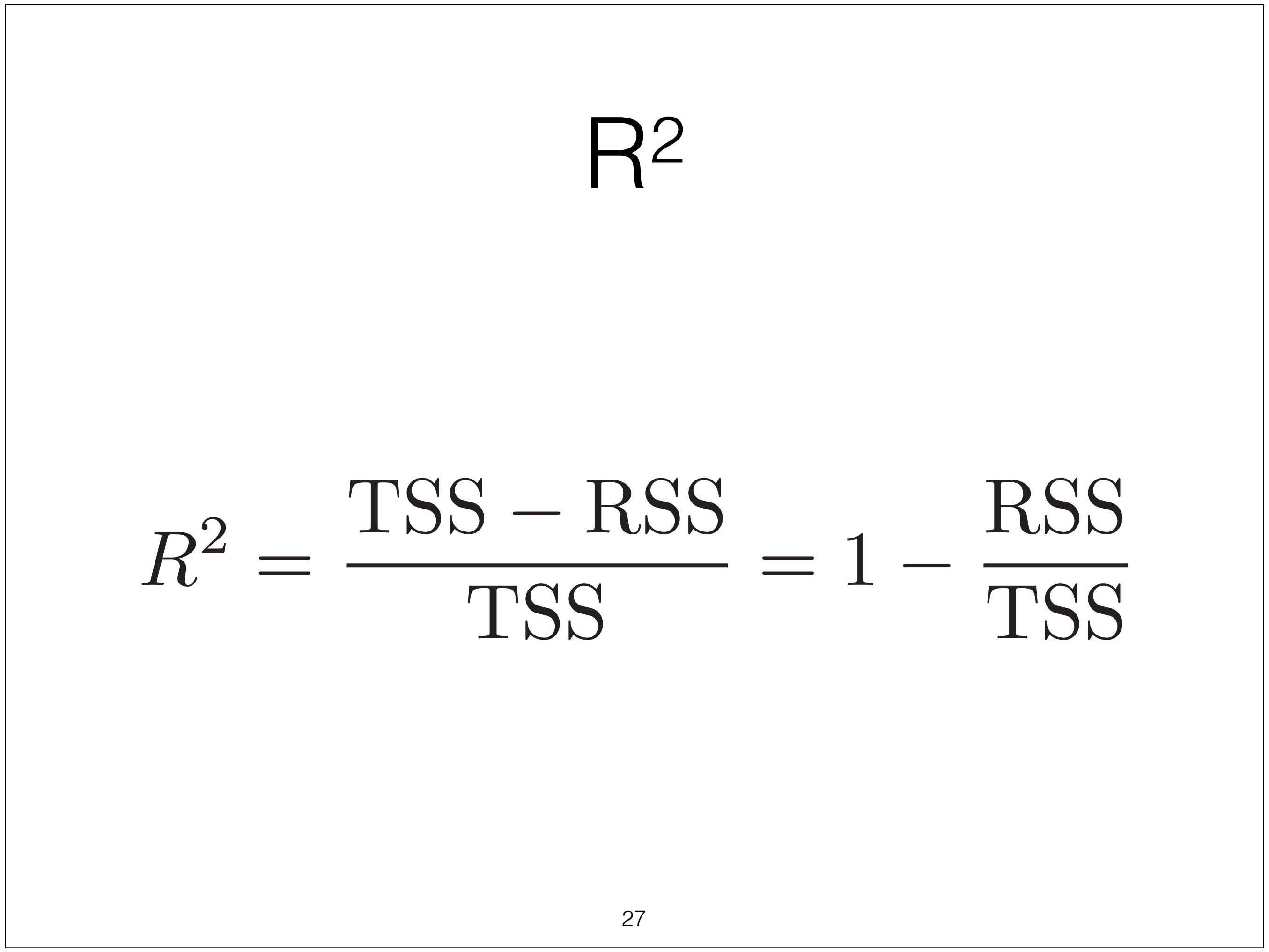


Overall Accuracy: RSE

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

Once we have rejected the null hypothesis, it is natural to ask to which extent the model fits the data? The residual standard error or RSE and the R statistic are typically used to assess the linear regression fit.

RSE, roughly speaking, is the average amount that the response will deviate from the true regression. For the advertising data RSE is 3.26, i.e. the true regression line will deviate 3,260 units on average. Another way to think about this is, if the model were correct and true value of the coefficients were known, any prediction on sales based on TV will be off by about 3,260 units. In terms of the overall error for this data, well in total there are 14,000 units, so the error is 23%. So a high RSE indicates the model does not fit well.



R^2

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

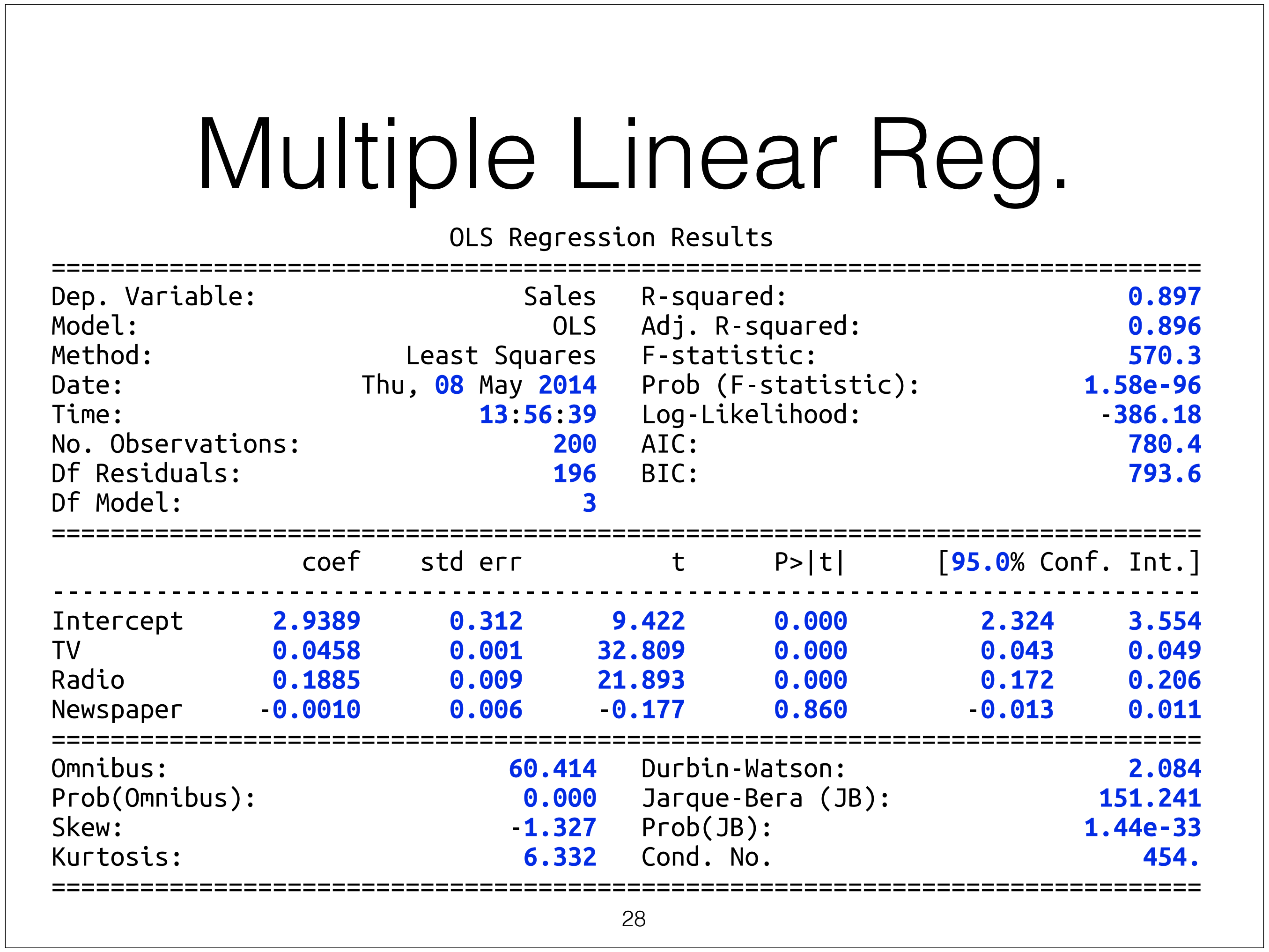
RSE is measured in the units of Y, and it is not always clear what a good RSE should be. The R provides an alternative fit as a proportion: namely the proportion of variance explained and takes a value from 0 to 1 and is independent of the scale of Y.

Here TSS is the total sum of squares (y_0-mean(Y)) + (y_1 - mean(Y)) + …TSS can be though of as the amount of variability inherent in the response before regression was performed. RSS is the measure of variability that is left unexplained after regression. So TSS-RSS is the amount of variability in the response that is explained (or removed) by regression.

A R^2 close to 1 indicates a large proportion of variability is explained in the response. A value near 0 means that the regression did not explain much of the variability in the response, which might happen if the linear model is wrong, or the inherent error is too high or both!

For our model we have an R^2 of 0.612, so just 2/3 of the variability in sales is explained by a linear regression on TV.

Note in physics, we tend to see values close to 1 and smaller values might indicate a problem. In biology, psychology, marketing etc. where the linear model is a rough estimate of the model, we see smaller values.



Multiple Linear Reg.

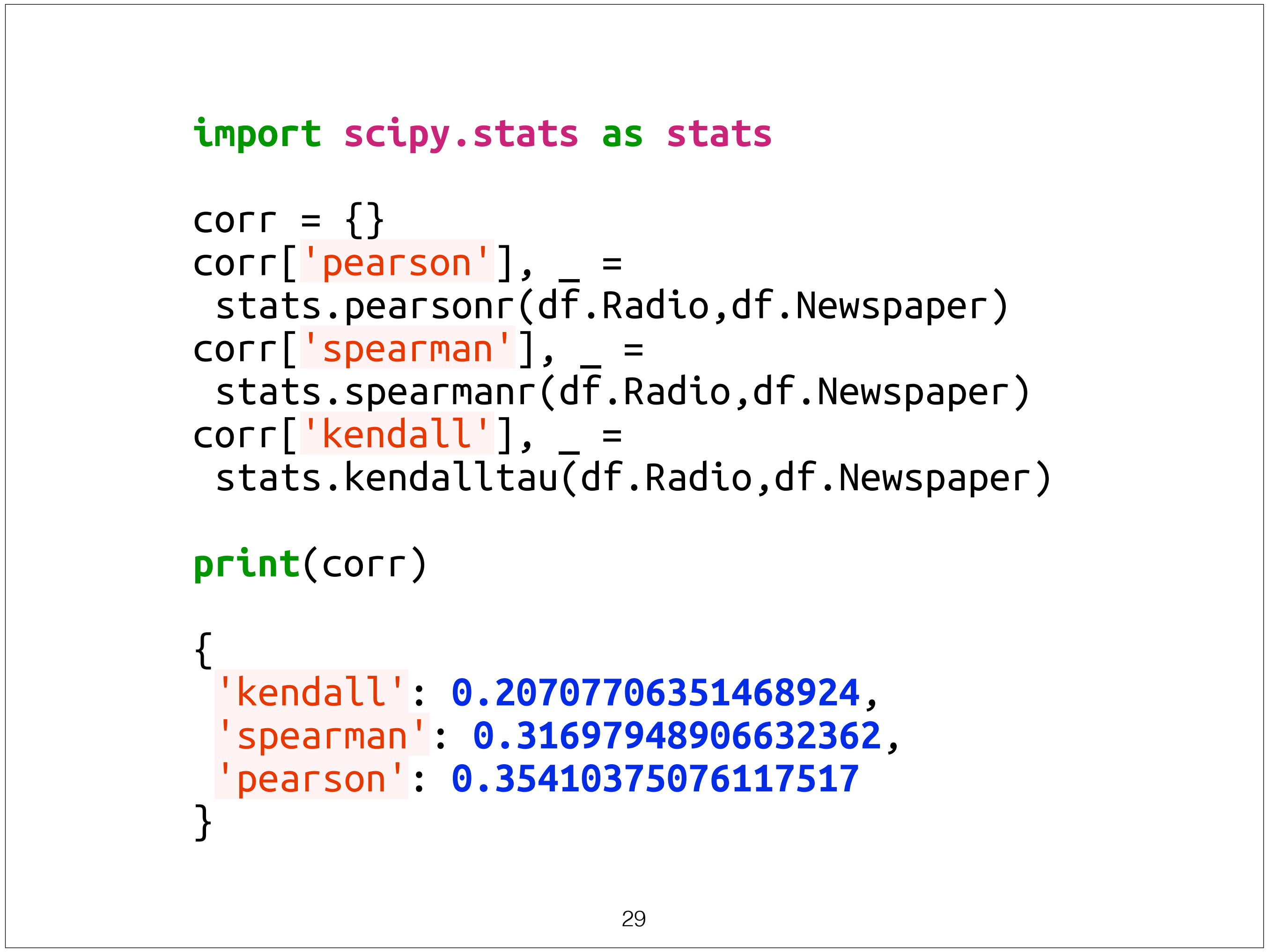
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.897
Model:                            OLS   Adj. R-squared:                  0.896
Method:                 Least Squares   F-statistic:                     570.3
Date:                Thu, 08 May 2014   Prob (F-statistic):           1.58e-96
Time:                        13:56:39   Log-Likelihood:                -386.18
No. Observations:                 200   AIC:                             780.4
Df Residuals:                     196   BIC:                             793.6
Df Model:                           3
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      2.9389      0.312      9.422      0.000         2.324     3.554
TV             0.0458      0.001     32.809      0.000         0.043     0.049
Radio          0.1885      0.009     21.893      0.000         0.172     0.206
Newspaper     -0.0010      0.006     -0.177      0.860        -0.013     0.011
==============================================================================
Omnibus:                       60.414   Durbin-Watson:                   2.084
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              151.241
Skew:                          -1.327   Prob(JB):                     1.44e-33
Kurtosis:                       6.332   Cond. No.                         454.
==============================================================================
```

Instead of having multiple linear models for each feature, we could have a single model with multiple predictors. We can interpret the coefficients as the average effect on Y for a unit increase in the particular feature while holding all other predictors fixed. For e.g. for sales=$\beta_0 + \beta_1 \times$TV+$\beta_2 \times$radio+$\beta_3 \times$newspaper+ε.

Geometrically we are fitting a hyper-plane with the specific slopes given by the least square regression. There is an issue about correlated predictors here. What happens if two of the predictors are highly correlated? Then the variance of the coefficients increases. Interpretation also suffers, since we can't claim to keep one predictor constant since they will both move together.

So claims of causality should be avoided for observational data, since in reality predictors change together.

Here we see by a t statistic of -0.177 that is does not have an effect, or the p-value is close to 1 implying the null hypothesis. But we have to be careful how to interpret this. On it's own newspaper could be a good predictor, but in the presence of TV and radio, newspaper is not showing significance.



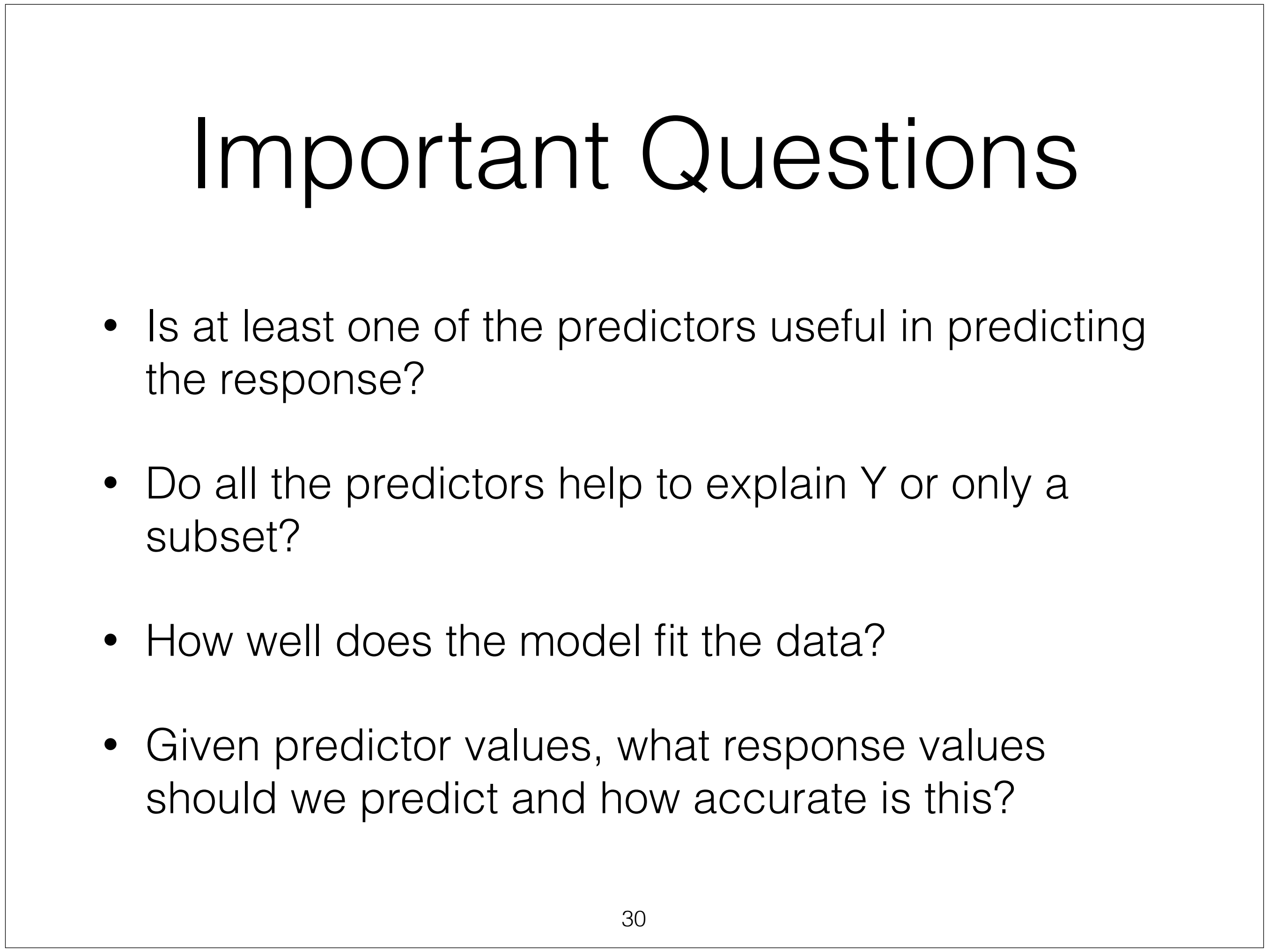
```
import scipy.stats as stats

corr = {}
corr['pearson'], _ =
 stats.pearsonr(df.Radio,df.Newspaper)
corr['spearman'], _ =
 stats.spearmanr(df.Radio,df.Newspaper)
corr['kendall'], _ =
 stats.kendalltau(df.Radio,df.Newspaper)

print(corr)

{
 'kendall': 0.20707706351468924,
 'spearman': 0.31697948906632362,
 'pearson': 0.35410375076117517
}
```

An in fact if we look at the correlations of the Radio and Newspaper, we see that they are correlated. So any effect of newspaper has been soaked up the Radio since they are correlated. So with Radio in the model newspaper is no longer needed. It doesn't tell us anything more, or improve the prediction, given we have measured the Radio data correctly.



Important Questions

- Is at least one of the predictors useful in predicting the response?
- Do all the predictors help to explain Y or only a subset?
- How well does the model fit the data?
- Given predictor values, what response values should we predict and how accurate is this?

The first three we can answer by looking at the model. The last question we will get to later, but we can use confidence intervals to quantify the uncertainty around the predictor.