A Python Companion to ISLR

Naresh Gurbuxani

June 4, 2019

Contents

1 Introduction

Figure 1 shows graphs of Wage versus three variables.

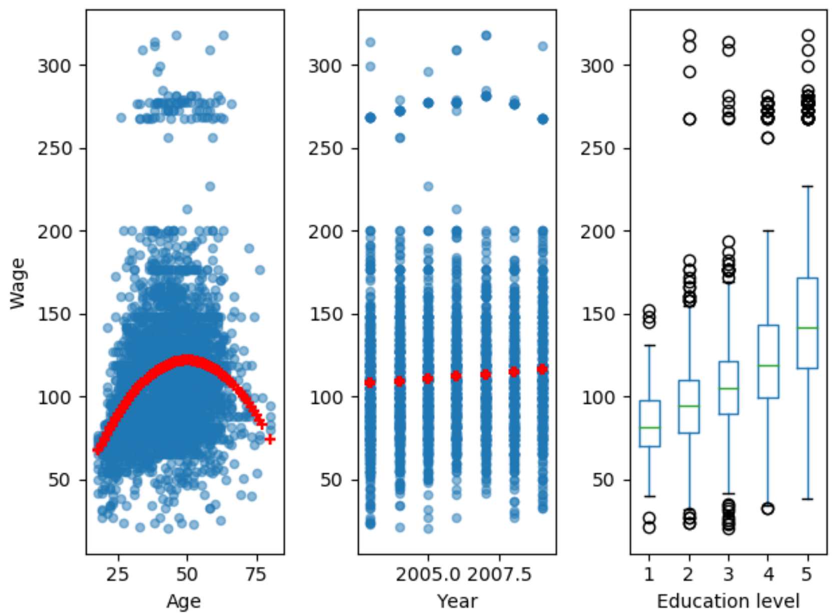

Figure 1: `Wage` data, which contains income survey information for males from the central Atlantic region of the United States. Left: `wage` as a function of `age`. On average, `wage` increases with `age` until about 60 years of age, at which point it begins to decline. Center: `wage` as a function of `year`. There is a slow but steady increase of approximately \$10,000 in the average `wage` between 2003 and 2009. Right: Boxplots displaying `wage` as a function of `education`, with 1 indicating the lowest level (no highschool diploma) and 5 the highest level (an advanced graduate degree). On average, `wage` increases with the level of `education`.

Figure 2 shows boxplots of previous days' percentage changes in S&P

500 grouped according to today's change `Up` or `Down`.

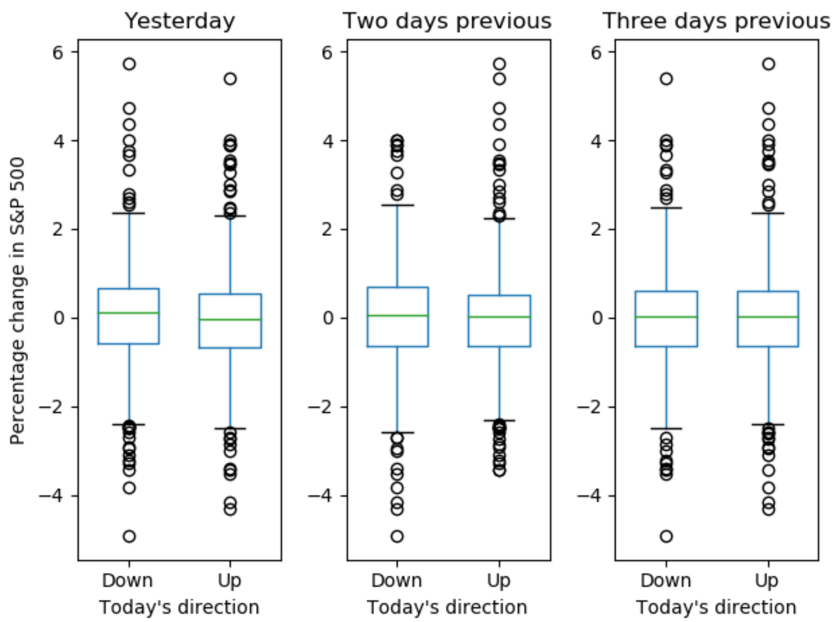

Figure 2: Left: Boxplots of the previous day's percentage change in the S&P 500 index for the days for which the market increased or decreased, obtained from the `Smarket` data. Center and Right: Same as left panel, but the percentage changes for two and three days previous are shown.

2 Statistical Learning

2.1 What is Statistical Learning?

Figure 3 shows scatter plots of `sales` versus `TV`, `radio`, and `newspaper` advertising. In each panel, the figure also includes an OLS regression line.

Figure 4 is a plot of `Income` versus `Years of Education` from the Income data set. In the left panel, the "true" function (given by blue line) is actually my guess.

Figure 5 is a plot of `Income` versus `Years of Education` and `Seniority` from the `Income` data set. Since the book does not provide the true values of `Income`, "true" values shown in the plot are actually third order polynomial fit.

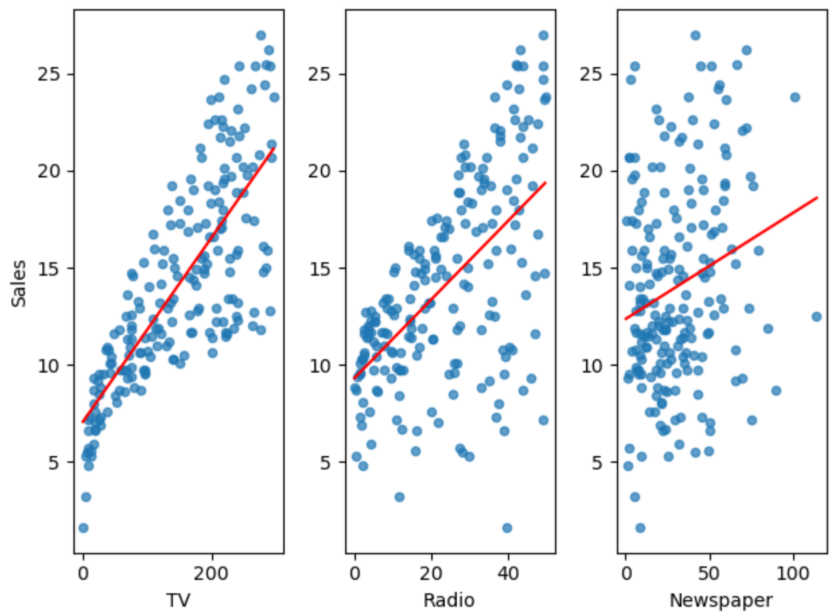

Figure 3: The `Advertising` data set. The plot displays `sales`, in thousands of units, as a function of `TV`, `radio`, and `newspaper` budgets, in thousands of dollars, for 200 different markets. In each plot we show the simple least squares fit of `sales` to that variable. In other words, each red line represents a simple model that can be used to predict `sales` using `TV`, `radio`, and `newspaper`, respectively.

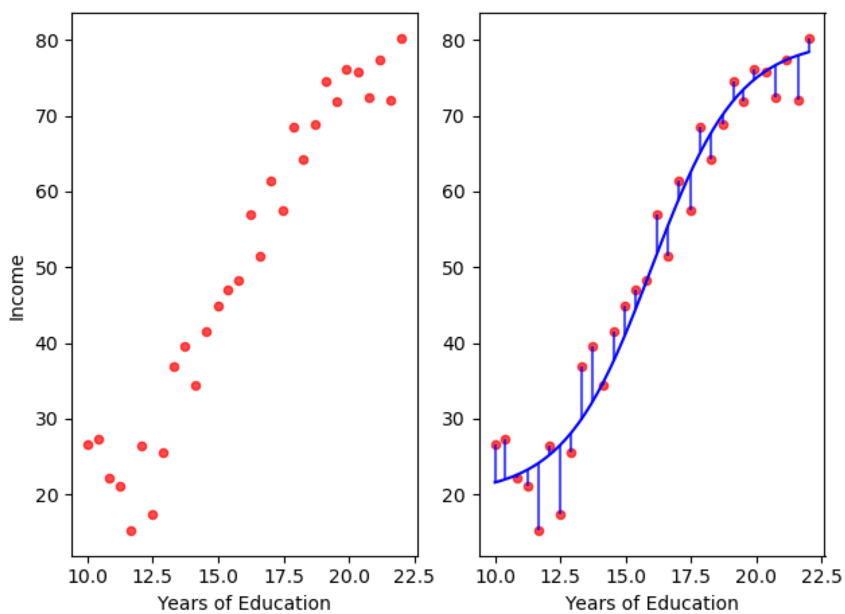

Figure 4: The `Income` data set. Left: The red dots are the observed values of `income` (in tens of thousands of dollars) and `years of education` for 30 individuals. Right: The blue curve represents the true underlying relationship between `income` and `years of education`, which is generally unknown (but is known in this case because the data are simulated). The vertical lines represent the error associated with each observation. Note that some of the errors are positive (when an observation lies above the blue curve) and some are negative (when an observation lies below the curve). Overall, these errors have approximately mean zero.

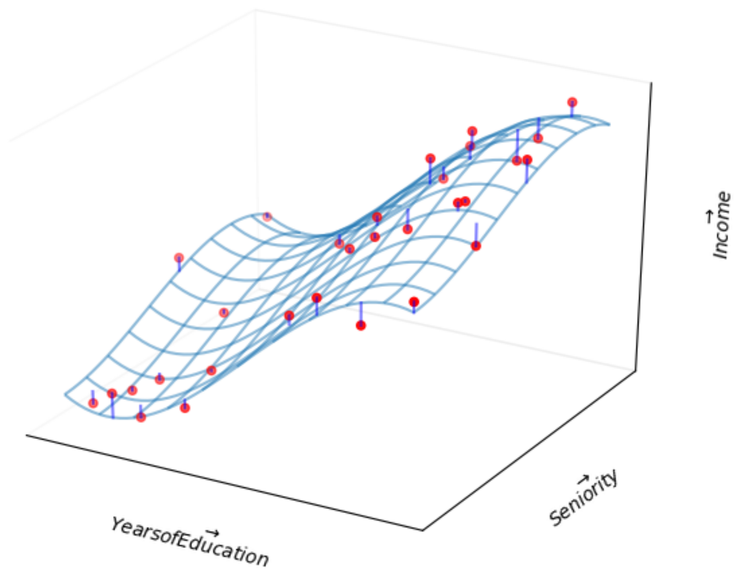

Figure 5: The plot displays `income` as a function of `years of education` and `seniority` in the `Income` data set. The blue surface represents the true underlying relationship between `income` and `years of education` and `seniority`, which is known since the data are simulated. The red dots indicate the observed values of these quantities for 30 individuals.

Figure 6 shows an example of the parametric approach applied to the `Income` data from previous figure.

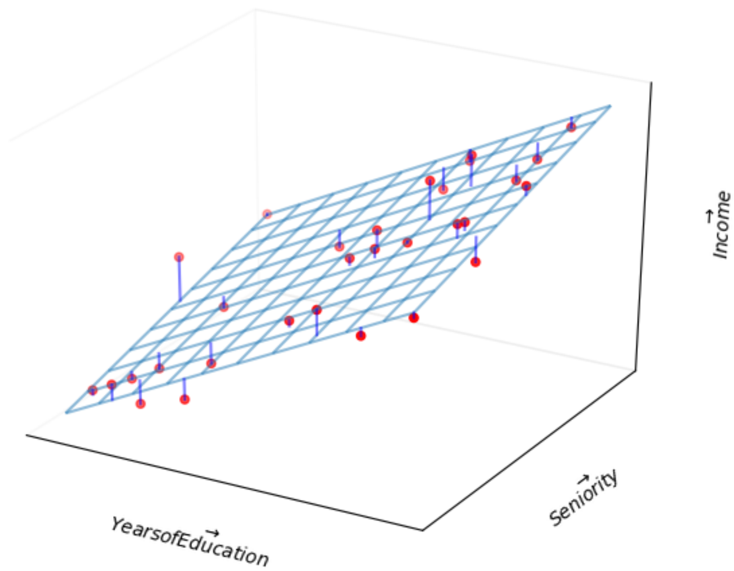

Figure 6: A linear model fit by least squares to the `Income` data from figure 5. The observations are shown in red, and the blue plane indicates the least squares fit to the data.

Figure 7 provides an illustration of the trade-off between flexibility and interpretability for some of the methods covered in this book.

Figure 8 provides a simple illustration of the clustering problem.

2.2 Assessing Model Accuracy

Figure 9 illustrates the tradeoff between training MSE and test MSE. We select a "true function" whose shape is similar to that shown in the book. In the left panel, the orange, blue, and green curves illustrate three possible estimates for f given by the black curve. The orange line is the linear regression fit, which is relatively inflexible. The blue and green curves were produced using *smoothing splines* from `UnivariateSpline` function in `scipy` package. We obtain different levels of flexibility by varying the parameter `s`, which affects the number of knots.

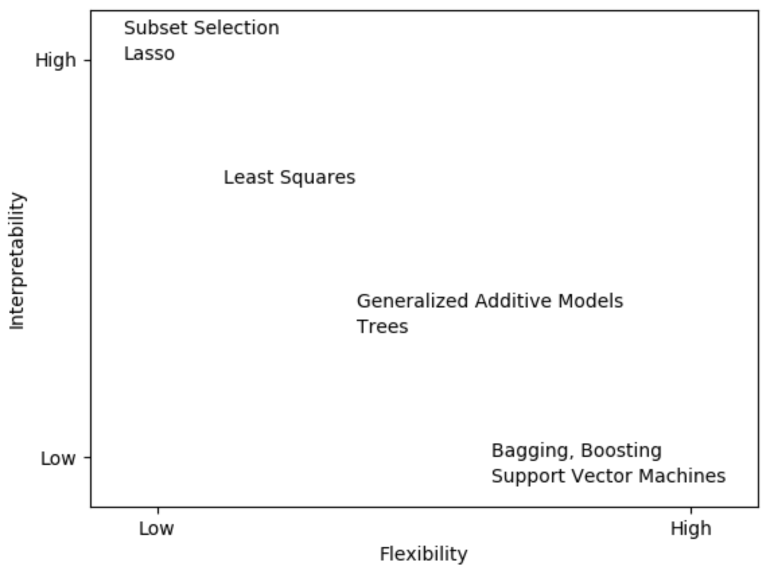

Figure 7: A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.

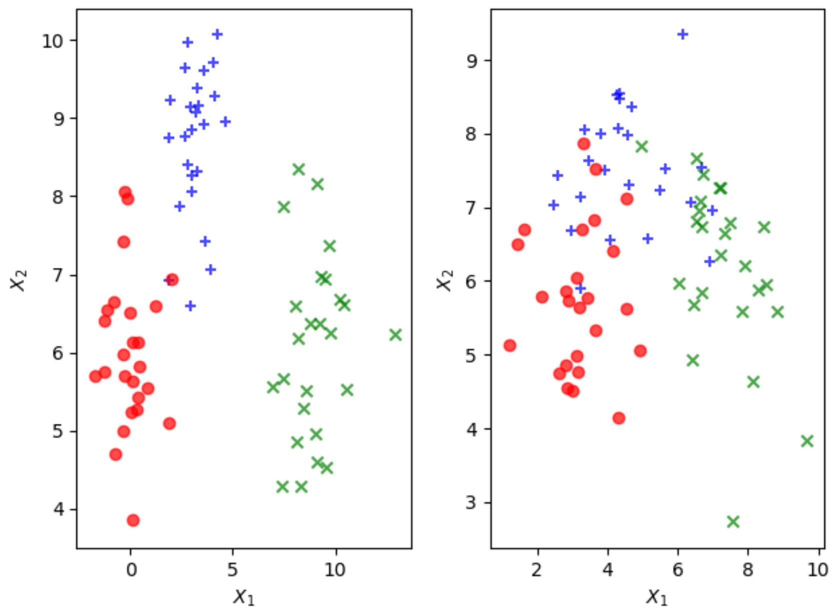

Figure 8: A clustering data set involving three groups. Each group is shown using a different colored symbol. Left: The three groups are well-separated. In this setting, a clustering approach should successfully identify the three groups. Right: There is some overlap among the groups. Now the clustering taks is more challenging.

For the right panel, we have chosen polynomial fits. The degree of polynomial represents the level of flexibility. This is because the function `UnivariateSpline` does not more than five degrees of freedom.

When we repeat the simulations for figure 9, we see considerable variation in the right panel MSE plots. But the overall conclusion remains the same.

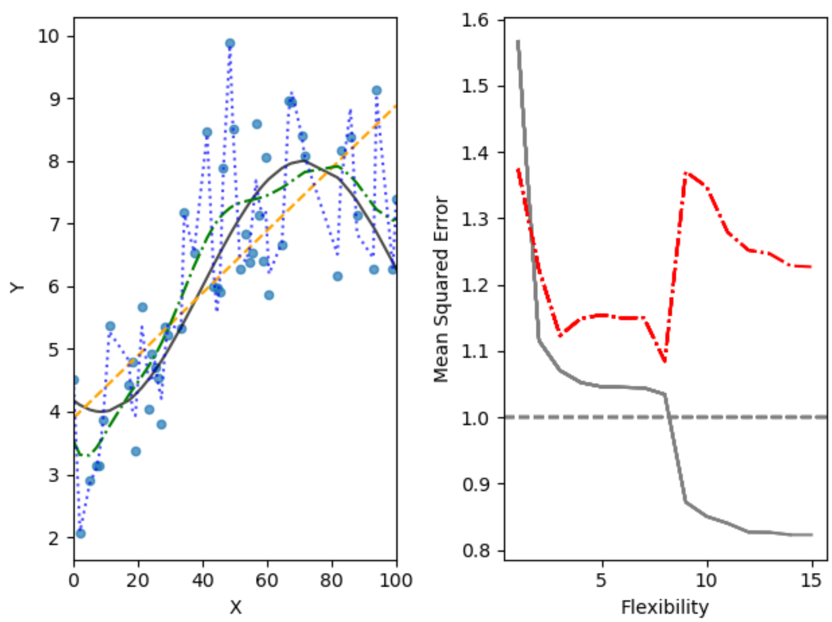

Figure 9: Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed grey line).

Figure 10 provides another example in which the true f is approximately linear.

Figure 11 displays an example in which f is highly non-linear. The training and test MSE curves still exhibit the same general patterns.

Figure 12 displays the relationship between bias, variance, and test MSE. This relationship is referred to as *bias-variance trade-off*. When simulations are repeated, we see considerable variation in different graphs, especially for MSE lines. But overall shape remains the same.

Figure 13 provides an example using a simulated data set in two-dimensional

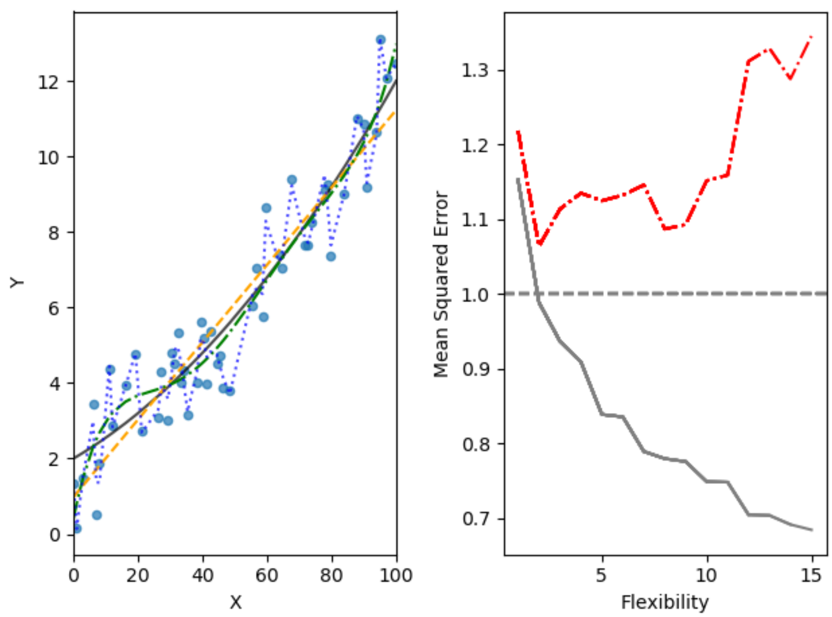

Figure 10: Details are as in figure 9 using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data.

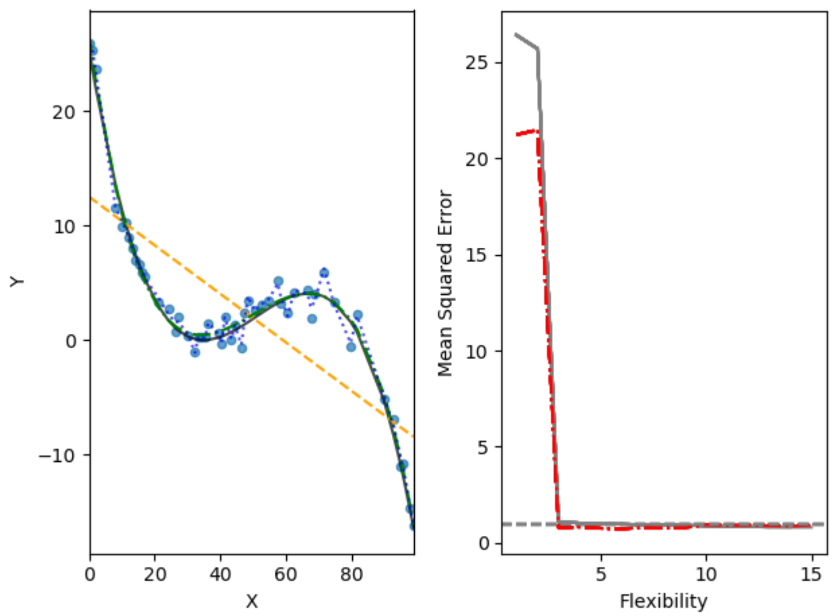

Figure 11: Details are as in figure 9, using a different f that is far from linear. In this setting, linear regression provides a very poor fit to the data.

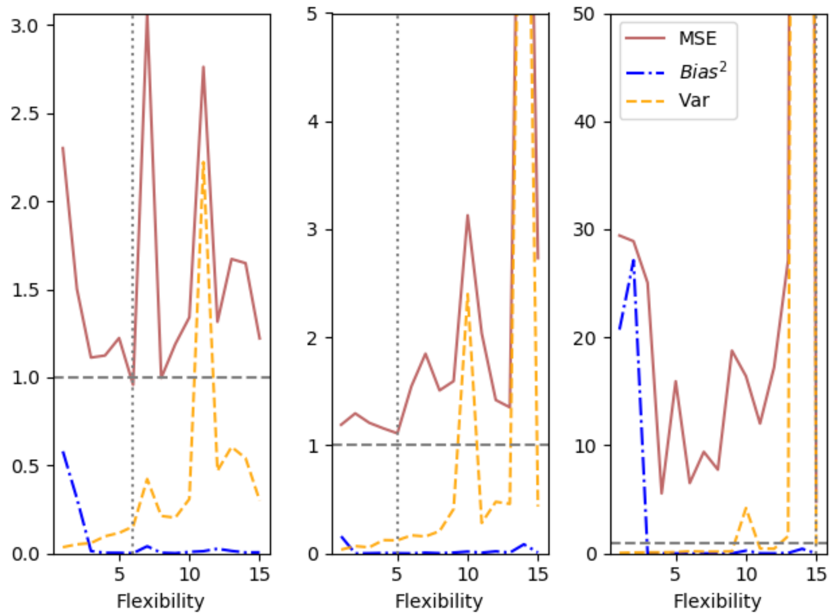

Figure 12: Squared bias (blue curve), variance (orange curve), $Var(\epsilon)$ (dashed line), and test MSE (red curve) for the three data sets in figures 9 - 11. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.

space consisting of predictors X_1 and X_2.

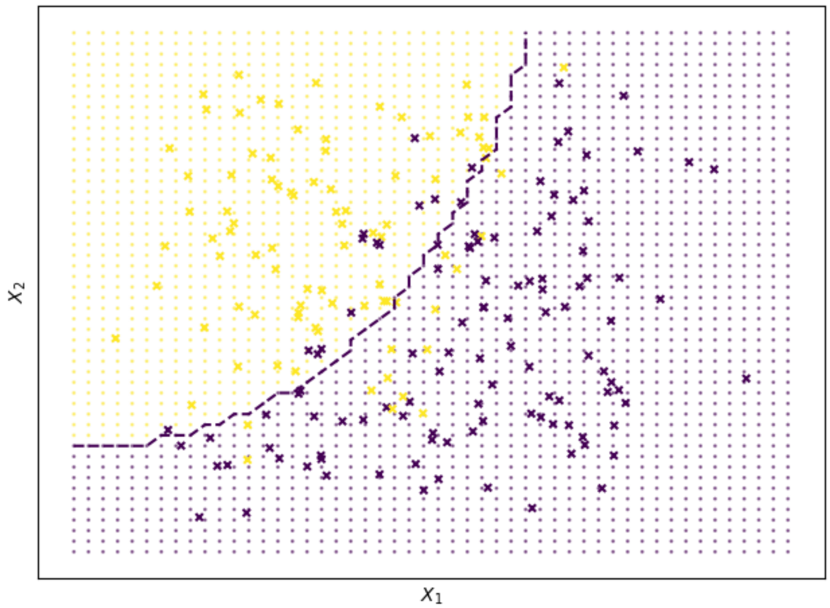

Figure 13: A simulated data set consisting of 200 observations in two groups, indicated in blue and orange. The dashed line represents the Bayes decision boundary. The orange background grid indicates the region in which a test observation will be assigned to the orange class, and blue background grid indicates the region in which a test observation will be assigned to the blue class.

Figure 14 displays the KNN decision boundary, using $K = 10$, when applied to the simulated data set from figure 13. Even though the true distribution is not known by the KNN classifier, the KNN decision making boundary is very close to that of the Bayes classifier.

In figure 16 we have plotted the KNN test and training errors as a function of $\frac{1}{K}$. As $\frac{1}{K}$ increases, the method becomes more flexible. As in the regression setting, the training error rate consistently declines as the flexibility increases. However, the test error exhibits the characteristic U-shape, declining at first (with a minimum at approximately $K = 10$) before increasing again when the method becomes excessively flexible and overfits.

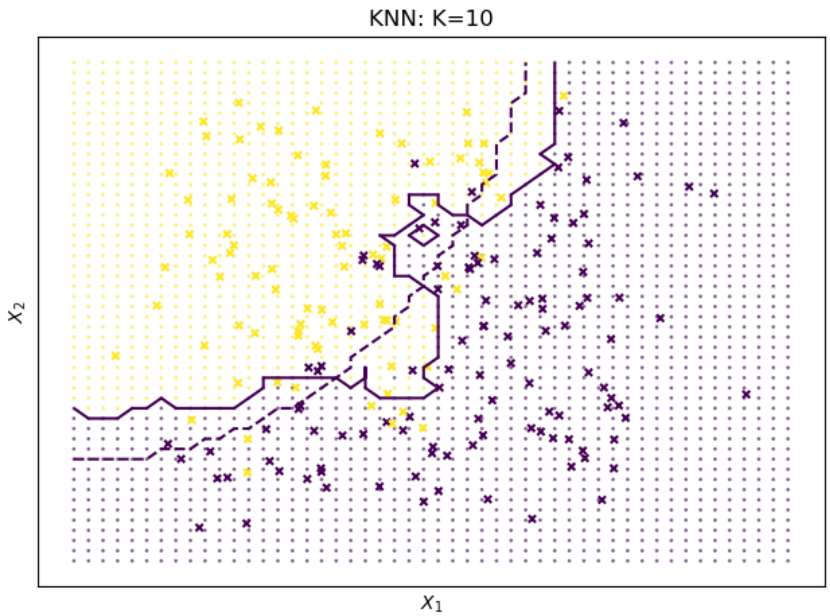

Figure 14: The firm line indicates the KNN decision boundary on the data from figure 13, using $K = 10$. The Bayes decision boundary is shown as a dashed line. The KNN and Bayes decision boundaries are very similar.

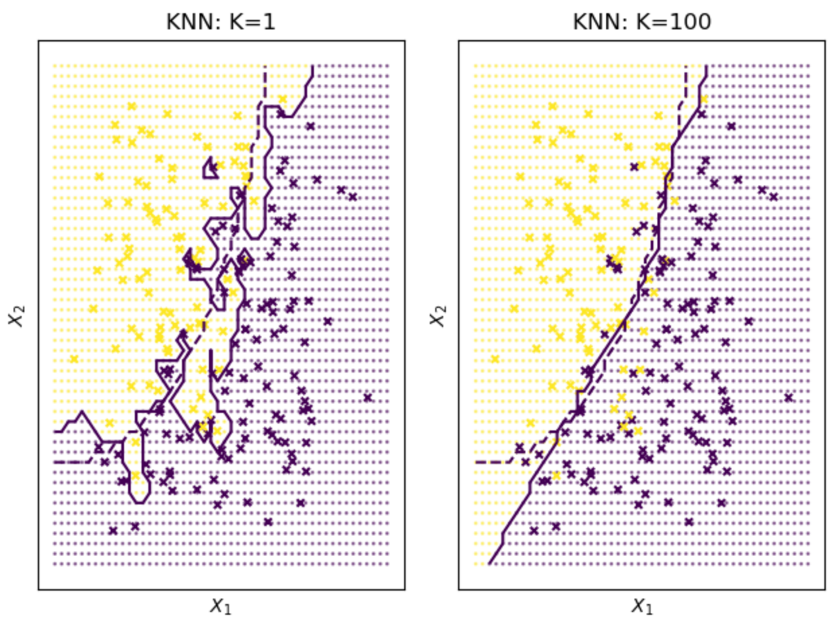

Figure 15: A comparison of the KNN decision boundaries (solid curves) obtained using $K = 1$ and $K = 100$ on the data from figure 13. With $K = 1$, the decision boundary is overly flexible, while with $K = 100$ it is not sufficiently flexible. The Bayes decision boundary is shown as dashed line.

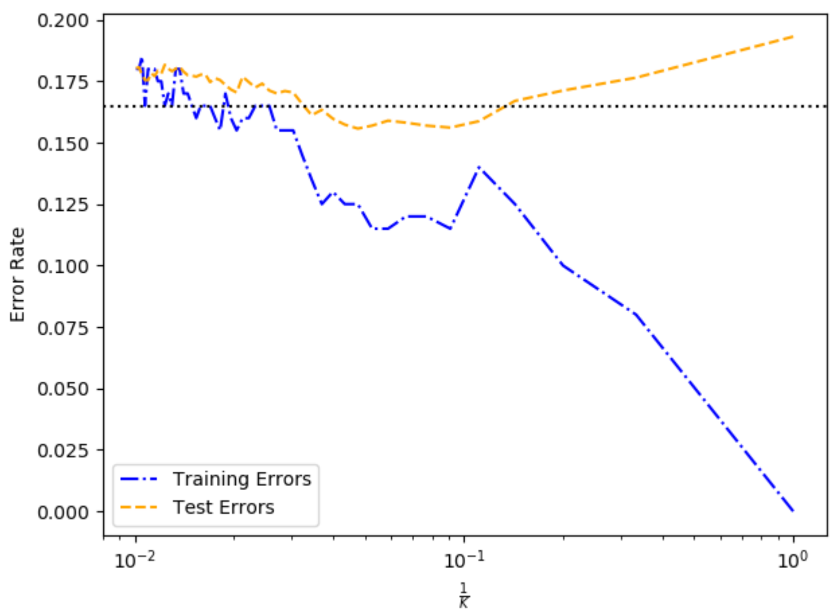

Figure 16: The KNN training error rate (blue, 200 observations) and test error rate (orange, 5,000 observations) on the data from figure 13 as the level of flexibility (assessed using $\frac{1}{K}$) increases, or equivalently as the number of neighbors K decreases. The black dashed line indicates the Bayes error rate.

2.3 Lab: Introduction to Python

2.3.1 Basic Commands

In Python a list can be created by enclosing comma-separated elements by square brackets. Length of a list can be obtained using len function.

```
x = [1, 3, 2, 5]
print(len(x))
y = 3
z = 5
print(y + z)
```

```
4
8
```

To create an array of numbers, use array function in numpy library. numpy functions can be used to perform element-wise operations on arrays.

```
import numpy as np
x = np.array([[1, 2], [3, 4]])
y = np.array([6, 7, 8, 9]).reshape((2, 2))
print(x)
print(y)
print(x ** 2)
print(np.sqrt(y))
```

```
[[1 2]
 [3 4]]
[[6 7]
 [8 9]]
[[ 1  4]
 [ 9 16]]
[[2.44948974 2.64575131]
 [2.82842712 3.        ]]
```

numpy.random has a number of functions to generate random variables that follow a given distribution. Here we create two correlated sets of numbers, x and y, and use numpy.corrcoef to calculate correlation between them.

```
import numpy as np
np.random.seed(911)
x = np.random.normal(size=50)
y = x + np.random.normal(loc=50, scale=0.1, size=50)
print(np.corrcoef(x, y))
print(np.corrcoef(x, y)[0, 1])
print(np.mean(x))
print(np.var(y))
print(np.std(y) ** 2)
```

```
[[1.         0.99374931]
 [0.99374931 1.        ]]
0.9937493134584551
-0.020219724397254404
0.9330621750073689
0.9330621750073688
```

2.3.2 Graphics

matplotlib library has a number of functions to plot data in Python. It is possible to view graphs on screen or save them in file for inclusion in a document.

```
import numpy as np
import matplotlib              # only if we need to save figure in file
matplotlib.use('Agg')          # only to save figure in file
import matplotlib.pyplot as plt

x = np.random.normal(size=100)
y = np.random.normal(size=100)
plt.plot(x, y)
plt.xlabel('This is x-axis')
plt.ylabel('This is y-axis')
plt.title('Plot of X vs Y')

plt.savefig('xyPlot.png')      # only to save figure in a file
```

numpy function linspace can be used to create a sequence between a start and an end of a given length.

```
import numpy as np
```

```
import matplotlib.pyplot as plt

x = np.linspace(-np.pi, np.pi, num=50)
y = x
xx, yy = np.meshgrid(x, y)
zz = np.cos(yy) / (1 + xx ** 2)

plt.contour(xx, yy, zz)

fig, ax = plt.subplots()
zza = (zz - zz.T) / 2.0
CS = ax.contour(xx, yy, zza)
ax.clabel(CS, inline=1)
```

2.3.3 Indexing Data

To access elements of an array, specify indexes inside square brackets. It is possible to access multiple rows and columns. `shape` method gives number of rows followed by number of columns.

```
import numpy as np

A = np.array(np.arange(1, 17))
A = A.reshape(4, 4, order='F')  # column first, Fortran style
print(A)
print(A[1, 2])
print(A[(0,2),:][:,(1,3)])
print(A[range(0,3),:][:,range(1,4)])
print(A[range(0, 2), :])
print(A[:, range(0, 2)])
print(A[0,:])
print(A.shape)
```

```
[[ 1  5  9 13]
 [ 2  6 10 14]
 [ 3  7 11 15]
 [ 4  8 12 16]]
10
[ 5 15]
[ 5 10 15]
[[ 1  5  9 13]
```

```
 [ 2  6 10 14]]
[[1 5]
 [2 6]
 [3 7]
 [4 8]]
(4, 4)
```

2.3.4 Loading Data

`pandas` library provides `read_csv` function to read files with data in rectangular shape.

```
import pandas as pd
Auto = pd.read_csv('data/Auto.csv')
print(Auto.head())
print(Auto.shape)
print(Auto.columns)
```

```
    mpg  cylinders  displacement  ... year  origin                       name
0  18.0          8         307.0  ...   70       1  chevrolet chevelle malibu
1  15.0          8         350.0  ...   70       1          buick skylark 320
2  18.0          8         318.0  ...   70       1         plymouth satellite
3  16.0          8         304.0  ...   70       1              amc rebel sst
4  17.0          8         302.0  ...   70       1                ford torino

[5 rows x 9 columns]
(397, 9)
Index(['mpg', 'cylinders', 'displacement', 'horsepower', 'weight',
       'acceleration', 'year', 'origin', 'name'],
      dtype='object')
```

To load data from an `R` library, use `get_rdataset` function from `statsmodels`. This function seems to work only if the computer is connected to the internet.

```
from statsmodels import datasets
carseats = datasets.get_rdataset('Carseats', package='ISLR').data
print(carseats.shape)
print(carseats.columns)
```

```
(400, 11)
Index(['Sales', 'CompPrice', 'Income', 'Advertising', 'Population', 'Price',
       'ShelveLoc', 'Age', 'Education', 'Urban', 'US'],
      dtype='object')
```

2.3.5 Additional Graphical and Numerical Summaries

`plot` method can be directly applied to a `pandas` dataframe.

```
import pandas as pd
Auto = pd.read_csv('data/Auto.csv')
Auto.boxplot(column='mpg', by='cylinders', grid=False)
```

`hist` method can be applied to plot a histogram.

```
import pandas as pd
Auto = pd.read_csv('data/Auto.csv')
Auto.hist(column='mpg')
Auto.hist(column='mpg', color='red')
Auto.hist(column='mpg', color='red', bins=15)
```

For pairs plot, use `scatter_matrix` method in `pandas.plotting`.

```
import pandas as pd
from pandas import plotting
Auto = pd.read_csv('data/Auto.csv')
plotting.scatter_matrix(Auto[['mpg', 'displacement', 'horsepower', 'weight',
      'acceleration']])
```

On `pandas` dataframes, `describe` method produces a summary of each variable.

```
import pandas as pd
Auto = pd.read_csv('data/Auto.csv')
print(Auto.describe())
```

```
              mpg   cylinders  ...        year      origin
count  397.000000  397.000000  ...  397.000000  397.000000
mean    23.515869    5.458438  ...   75.994962    1.574307
std      7.825804    1.701577  ...    3.690005    0.802549
min      9.000000    3.000000  ...   70.000000    1.000000
25%     17.500000    4.000000  ...   73.000000    1.000000
50%     23.000000    4.000000  ...   76.000000    1.000000
75%     29.000000    8.000000  ...   79.000000    2.000000
max     46.600000    8.000000  ...   82.000000    3.000000

[8 rows x 7 columns]
```

3 Linear Regression

3.1 Simple Linear Regression

Figure 17 displays the simple linear regression fit to the `Advertising` data, where $\hat{\beta}_0 = 0.0475$ and $\hat{\beta}_1 = 7.0326$.

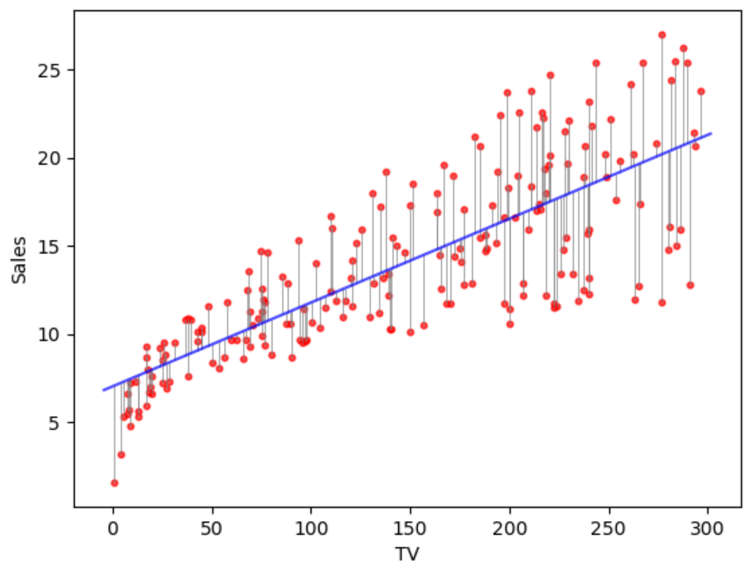

Figure 17: For the `Advertising` data, the least squares fit for the regression of `sales` onto `TV` is shown. The fit is found by minimizing the sum of squared errors. Each grey line represents an error, and the fit makes a compromise by averaging their squares. In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

In figure 18, we have computed RSS for a number of values of β_0 and β_1, using the advertising data with `sales` as the response and `TV` as the predictor.

The left-hand panel of figure 19 displays *population regression line* and *least squares line* for a simple simulated example. The red line in the left-hand panel displays the *true* relationship, $f(X) = 2 + 3X$, while the blue line is the least squares estimate based on observed data. In the right-hand panel of figure 19 we have generated five different data sets from the model $Y = 2 + 3X + \epsilon$ and plotted the corresponding five least squares lines.

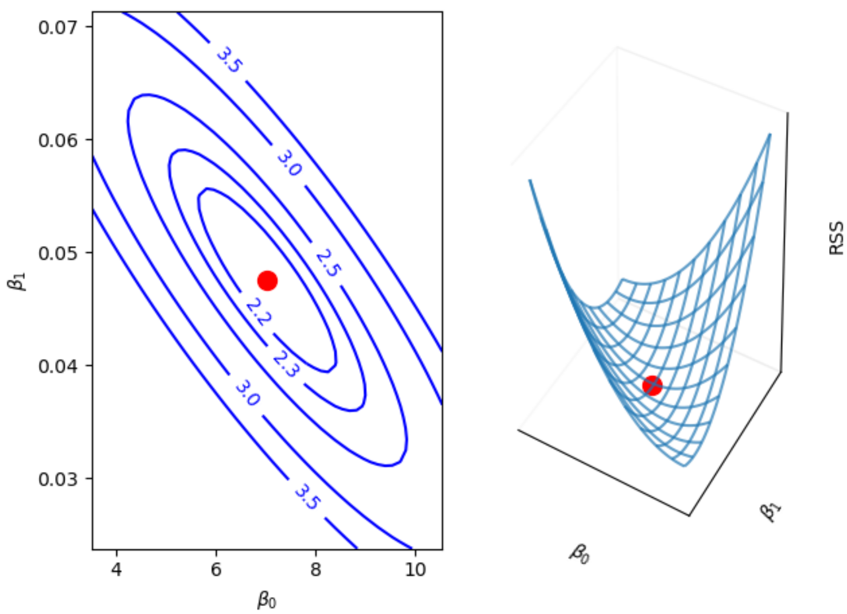

Figure 18: Contour and three-dimensional plots of the RSS on the `Advertising` data, using `sales` as the response and `TV` as the predictor. The red dots correspond to the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

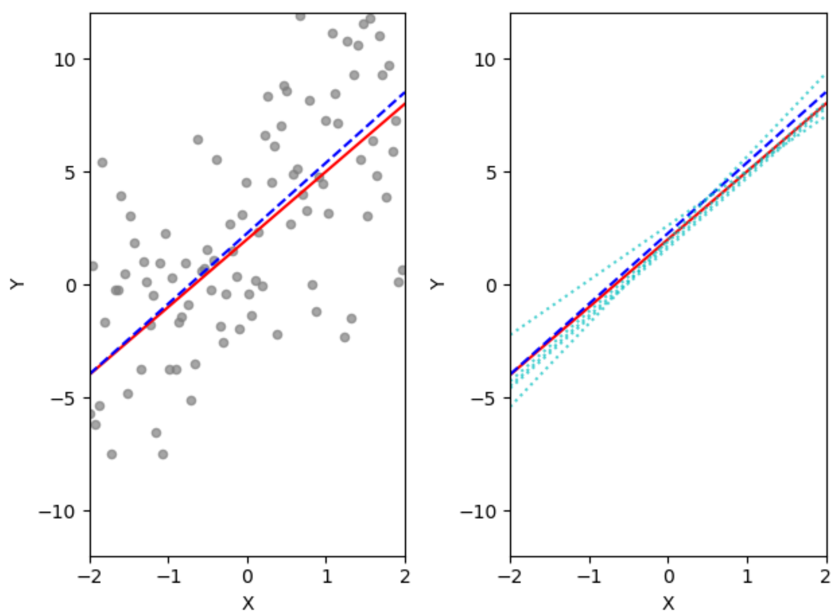

Figure 19: A simulated data set. Left: The red line represents the true relationship, $f(X) = 2 + 3X$, which is known as the population regression line. The blue line is the least squares line; it is the least squares estimate for $f(X)$ based on the observed data, shown in grey circles. Right: The population regression line is again shown in red, and the least squares line in blue. In cyan, five least squares lines are shown, each computed on the basis of a separate random set of observations. Each least squares line is different, but on average, the least squares lines are quite close to the population regression line.

For `Advertising` data, table 1 provides details of the least squares model for the regression of number of units sold on TV advertising budget.

	Coef.	Std.Err.	t	$P >\mid t \mid$
Intercept	7.0326	0.4578	15.3603	0.0
TV	0.0475	0.0027	17.6676	0.0

Table 1: For `Advertising` data, the coefficients of the least squares model for the regression of number of units sold on TV advertising budget. An increase of $1,000 on the TV advertising budget is associated with an increase in sales by around 50 units.

Next, in table 2, we report more information about the least squares model.

Quantity	Value
Residual standard error	3.259
R^2	0.612
F-statistic	312.145

Table 2: For the `Advertising` data, more information about the least squares model for the regression of number of units sold on TV advertising budget.

3.2 Multiple Linear Regression

Table 3 shows results of two simple linear regressions, each of which uses a different advertising medium as a predictor. We find that a $1,000 increase in spending on radio advertising is associated with an increase in sales by around 202 units. A $1,000 increase in advertising spending on on newspapers increases sales by approximately 55 units.

	Coef.	Std.Err.	t	$P >\mid t \mid$
Intercept	9.312	0.563	16.542	0.0
radio	0.202	0.02	9.921	0.0
Intercept	12.351	0.621	19.876	0.0
newspaper	0.055	0.017	3.3	0.001

Table 3: More simple linear regression models for `Advertising` data. Coefficients of the simple linear regression model for number of units sold on Top: radio advertising budget and Bottom: newspaper advertising budget. A $1,000 increase in spending on radio advertising is associated with an average increase sales by around 202 units, while the same increase in spending on newspaper advertising is associated with an average increase of around 55 units. `Sales` variable is in thousands of units, and the `radio` and `newspaper` variables are in thousands of dollars..

Figure 20 illustrates an example of the least squares fit to a toy data set with $p = 2$ predictors.

Table 4 displays multiple regression coefficient estimates when TV, radio, and newspaper advertising budgets are used to predict product sales using `Advertising` data.

	Coef.	Std.Err.	t	$P >\mid t \mid$
Intercept	2.939	0.312	9.422	0.0
TV	0.046	0.001	32.809	0.0
radio	0.189	0.009	21.893	0.0
newspaper	-0.001	0.006	-0.177	0.86

Table 4: For the `Advertising` data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.

Table 5 shows the correlation matrix for the three predictor variables and response variable in table 4.

Figure 21 displays a three-dimensional plot of `TV` and `radio` versus `sales`.

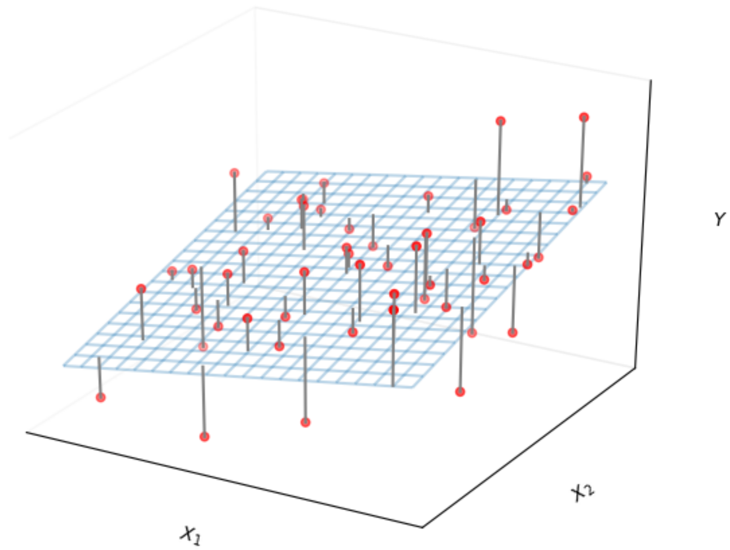

Figure 20: In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.

	TV	radio	newspaper	sales
TV	1.0	0.0548	0.0566	0.7822
radio	0.0548	1.0	0.3541	0.5762
newspaper	0.0566	0.3541	1.0	0.2283
sales	0.7822	0.5762	0.2283	1.0

Table 5: Correlation matrix for `TV`, `radio`, and `sales` for the `Advertising` data.

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570.0

Table 6: More information about the least squares model for the regression of number of units sold on TV, newspaper, and radio advertising budgets in the `Advertising` data. Other information about this model was displayed in table 4.

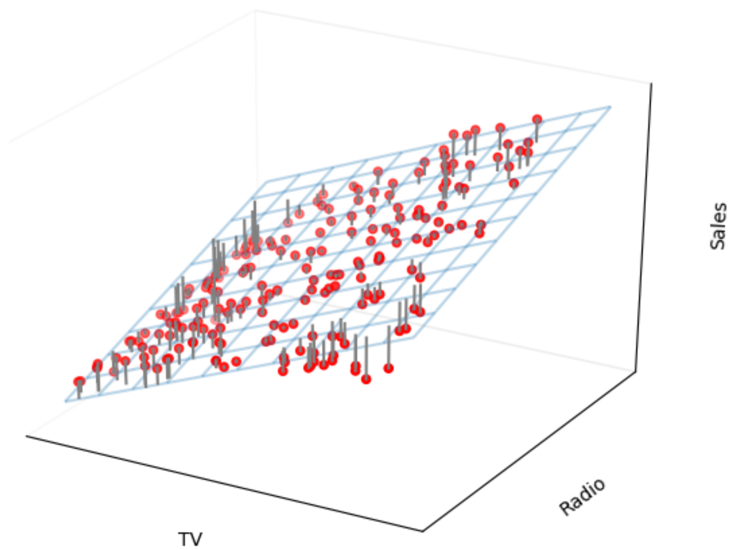

Figure 21: For the `Advertising` data, a linear regression fit to `sales` using `TV` and `radio` as predictors. From the pattern of the residuals, we can see that there is a pronounced non-linear relationship in the data. The positive residuals tend to lie along the 45-degree line, where TV and Radio budgets are split evenly. The negative residuals tend to lie away from this line, where budgets are more lopsided.

3.3 Other Considerations in the Regression Model

`Credit` data set displayed in figure 22 records `balance` (average credit card debt for a number of individuals) as well as several quantitative predictors: `age`, `cards` (number of credit cards), `education` and `rating` (credit rating).

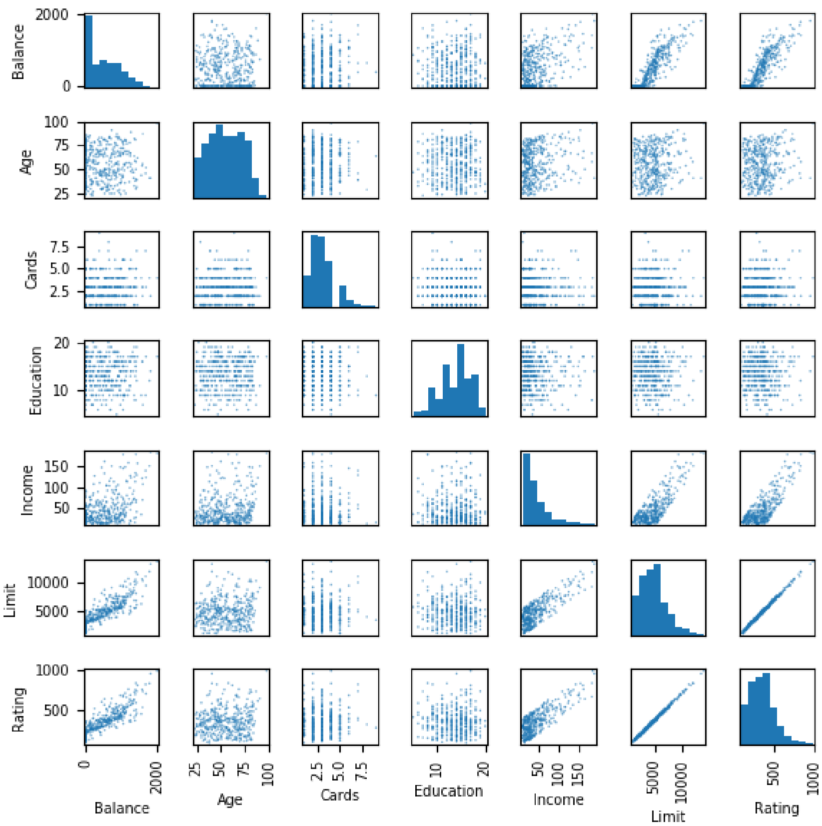

Figure 22: The `Credit` dataset contains information about `balance`, `age`, `cards`, `education`, `income`, `limit`, and `rating` for a number of potential customers.

Table 7 displays the coefficient estimates and other information associated with the model where `gender` is the only explanatory variable.

From table 8 we see that the estimated `balance` for the baseline, African American, is \$531.0. It is estimated that the Asian category will have an additional \$-18.7 debt, and that the Caucasian category will have an additional \$-12.5 debt compared to African American category.

	Coef.	Std.Err.	t	$P >\mid t \mid$
Intercept	509.803	33.128	15.389	0.0
Gender[T.Female]	19.733	46.051	0.429	0.669

Table 7: Least squares coefficient estimates associated with the regression of `balance` onto `gender` in the `Credit` data set.

	Coef.	Std.Err.	t	$P >\mid t \mid$
Intercept	531.0	46.319	11.464	0.0
Ethnicity[T.Asian]	-18.686	65.021	-0.287	0.774
Ethnicity[T.Caucasian]	-12.503	56.681	-0.221	0.826

Table 8: Least squares coefficient estimates associated with the regression of `balance` onto `ethnicity` in the `Credit` data set.

Table 9 shows results of regressing `sales` and `TV` and `radio` when an interaction term is included. Coefficient of interaction term `TV:radio` is highly significant.

In figure 23, the left panel shows least squares lines when we predict `balance` using `income` (quantitative) and `student` (qualitative variables). There is no interaction term between `income` and `student`. The right panel shows least squares lines when an interaction term is included.

	Coef.	Std.Err.	t	$P >\mid t \mid$
Intercept	6.75	0.248	27.233	0.0
TV	0.019	0.002	12.699	0.0
radio	0.029	0.009	3.241	0.001
TV:radio	0.001	0.0	20.727	0.0

Table 9: For `Advertising` data, least squares coefficient estimates associated with the regression of `sales` onto `TV` and `radio`, with an interaction term.

Figure 24 shows a scatter plot of `mpg` (gas mileage in miles per gallon) versus `horsepower` in the `Auto` data set. The figure also includes least squares fit line for linear, second degree, and fifth degree polynomials in `horsepower`.

Table 10 shows regression results of a quadratic fit to explain `mpg` as a function of `horsepower` and $horsepower^2$.

The left panel of figure 25 displays a residual plot from the linear regression of `mpg` onto `horsepower` on the `Auto` data set. The red line is a smooth fit to the residuals, which is displayed in order to make it easier to identify

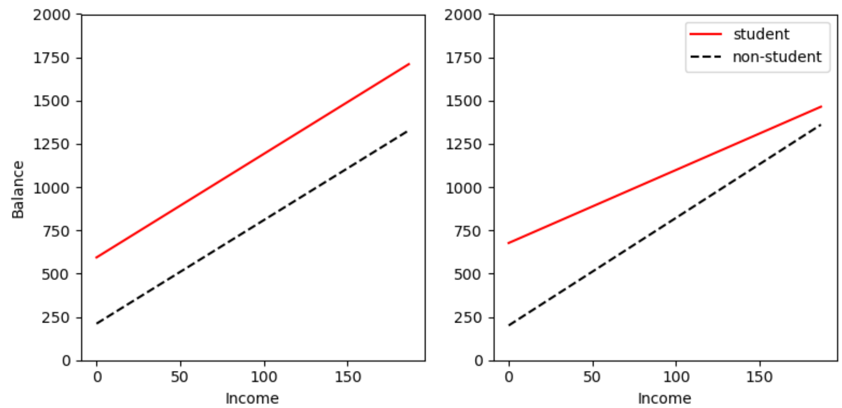

Figure 23: For the `Credit` data, the least squares lines are shown for prediction of `balance` from `income` for students and non-students. Left: There is no interaction between `income` and `student`. Right: There is an interaction term between `income` and `students`.

	Coef.	Std.Err.	t	$P >\mid t \mid$
Intercept	56.9001	1.8004	31.6037	0.0
horsepower	-0.4662	0.0311	-14.9782	0.0
$horsepower^2$	0.0012	0.0001	10.0801	0.0

Table 10: For the `Auto` data set, least squares coefficient estimates associated with the regression of `mpg` onto `horsepower` and

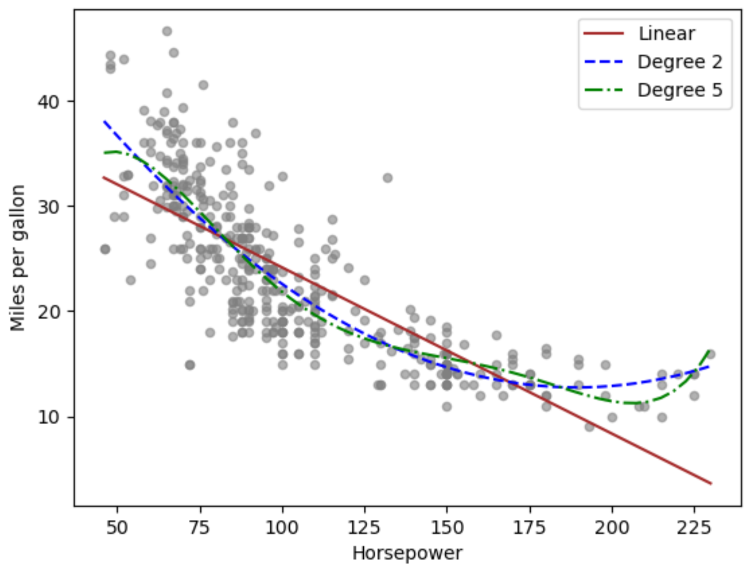

Figure 24: The `Auto` data set. For a number of cars, `mpg` and `horsepower` are shown. The linear regression fit is shown in orange. The linear regression fit for a model that includes first- and second-order terms of `horsepower` is shown as blue curve. The linear regression fit for a model that includes all polynomials of `horsepower` up to fifth-degree is shown in green.

any trends. The residuals exhibit a clear U-shape, which strongly suggests non-linearity in the data. In contrast, the right hand panel of figure25 displays the residual plot results from the model which contains a quadratic term in `horsepower`. Now there is little pattern in residuals, suggesting that the quadratic term improves the fit to the data.

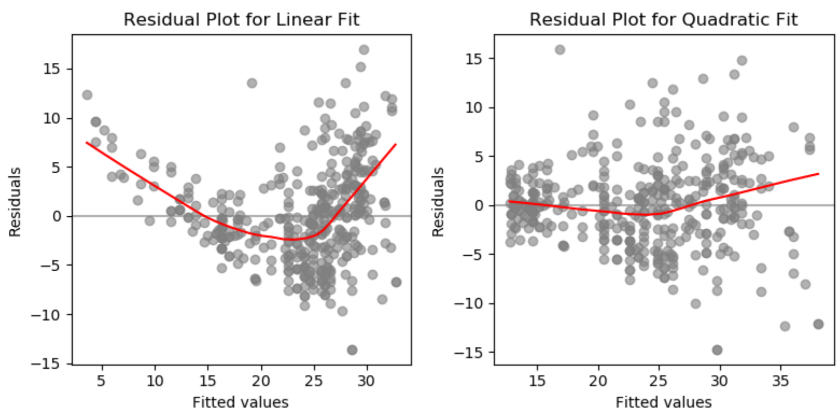

Figure 25: Plots of residuals versus predicted (or fitted) values for the `Auto` data set. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. Left: A linear regression of `mpg` on `horsepower`. A strong pattern in the residuals indicates non-linearity in the data. Right: A linear regression of `mpg` on `horsepower` and square of `horsepower`. Now there is little pattern in the residuals.

Figure 26 provides an illustration of correlations among residuals. In the top panel, we see the residuals from a linear regression fit to data generated with uncorrelated errors. There is no evidence of time-related trend in the residuals. In contrast, the residuals in the bottom panel are from a data set in which adjacent errors had a correlation of 0.9. Now there is a clear pattern in the residuals - adjacent residuals tend to take on similar values. Finally, the center panel illustrates a more moderate case in which the residuals had a correlation of 0.5. There is still evidence of tracking, but the pattern is less pronounced.

In the left-hand panel of figure 27, the magnitude of the residuals tends to increase with the fitted values. The right hand panel displays residual plot after transforming the response using $\log(Y)$. The residuals now appear to have constant variance, although there is some evidence of a non-linear relationship in the data.

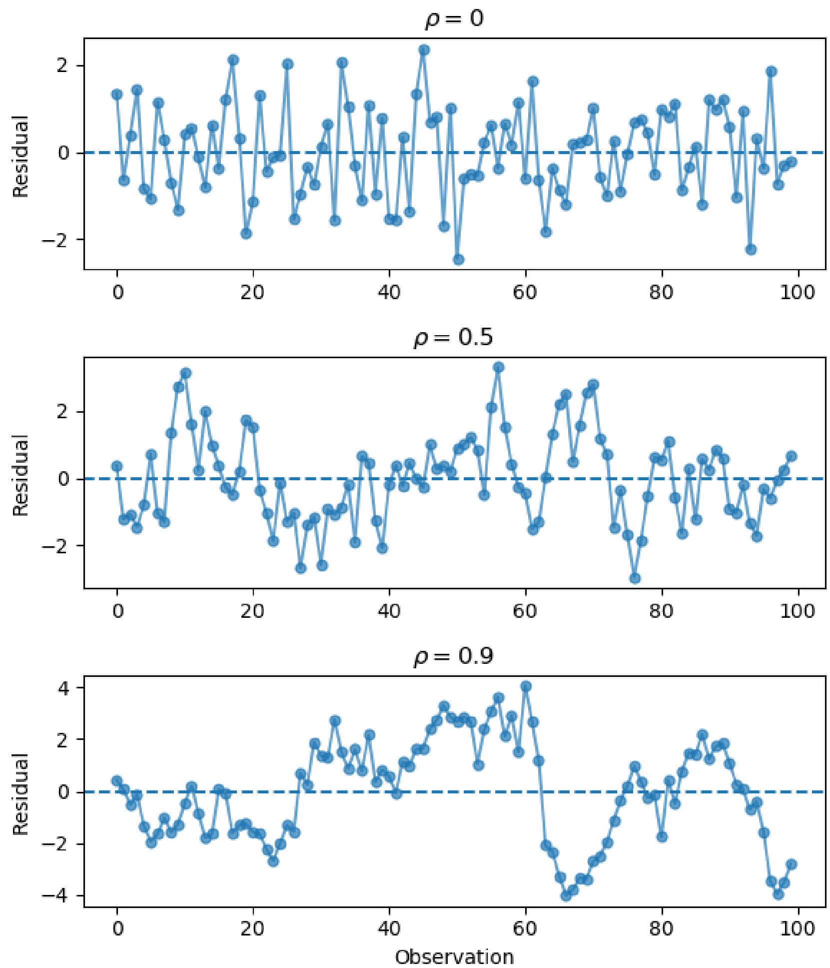

Figure 26: Plots of residuals from simulated time series data sets generated with differeing levels of correlation ρ between error terms for adjacent time points.

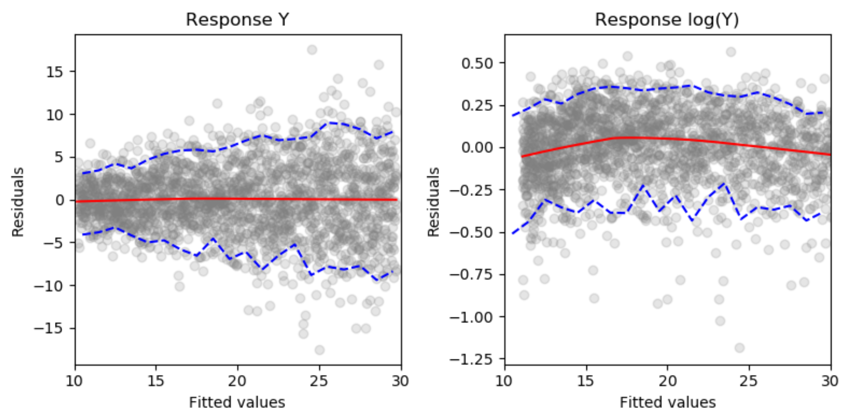

Figure 27: Residual plots. The red line, a smooth fit to the residuals, is intended to make it easier to identify a trend. The blue lines track 5^{th} and 95^{th} percentiles of the residuals, and emphasize patterns. Left: The funnel shape indicates heteroscedasticity. Right: the response has been log transformed, and now there is no evidence of heteroscedasticity.

The red point (observation 20) in the left hand panel of figure 28 illustrates a typical outlier. The red solid line is the least squares regression fit, while the blue dashed line is the least squares fit after removal of the outlier. In this case, removal of outlier has little effect on the least squares line. In the center panel of figure 28, the outlier is clearly visible. In practice, to decide if the outlier is sufficiently big to be considered an outlier, we can plot *studentized residuals*, computed by dividing each residual ϵ_i by its estimated standard error. These are shown in the right hand panel.

Observation 41 in the left-hand panel in figure 29 has high leverage, in that the predictor value for this observation is large relative to the other observations. The data displayed in figure 29 are the same as the data displayed in figure 28, except for the addition of a single high leverage observation[1]. The red solid line is the least squares fit to the data, while the blue dashed line is the fit produced when observation 41 is removed. Comparing the left-hand panels of figures 28 and 29, we observe that removing the high leverage observation has a much more substantial impact on least squares line than removing the outlier. The center panel of figure 29, for a data set with two predictors X_1 and X_2. While most of the observations' predictor values fall within the region of blue dashed lines, the red observation is well outside this

[1]The middle panel is from a different data set.

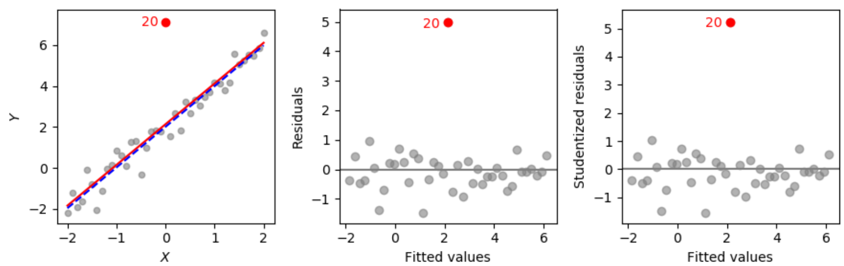

Figure 28: Left: The least squares regression line is shown in red. The regression line after removing the outlier is is shown in blue. Center: The residual plot clearly identifies the outlier. Right: The outlier has a studentized residual of 6; typically we expect values between -3 and 3.

range. But neither the value for X_1 nor the value for X_2 is unusual. So if we examine just X_1 or X_2, we will not notice this high leverage point. The right-panel of figure 29 provides a plot of studentized residuals versus h_i for the data in the left hand panel. Observation 41 stands out as having a very high leverage statistic as well as a high studentized residual.

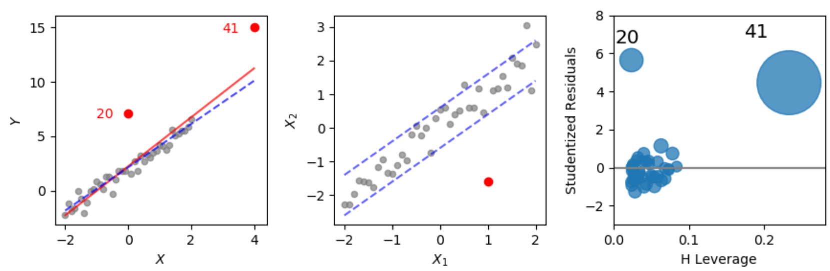

Figure 29: Left: Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the blue line is the fit with observation 41 removed. Center: The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and hence has high leverage. Right: Observation 41 has a high leverage and a high residual.

Figure 30 illustrates the concept of collinearity.

Figure 31 illustrates some of the difficulties that can result from collinearity. The left panel is a contour plot of the RSS associated with different possible coefficient estimates for the regression of `balance` on `limit` and `age`.

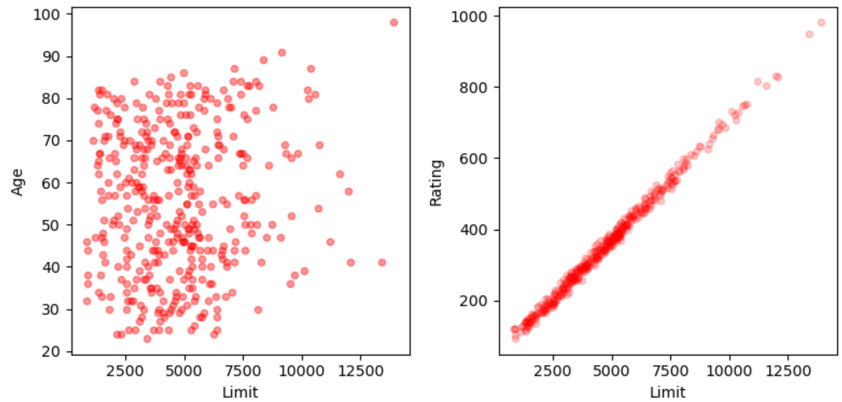

Figure 30: Scatter plots of the observations from the `Credit` data set. Left: A plot of `age` versus `limit`. These two variables not collinear. Right: A plot of `rating` versus `limit`. There is high collinearity.

Each ellipse represents a set of coefficients that correspond to the same RSS, with ellipses nearest to the center taking on the lowest values of RSS. The black dot and the associated dashed lines represent the coefficient estimates that result in the smallest possible RSS. The axes for `limit` and `age` have been scaled so that the plot includes possible coefficients that are up to four standard errors on either side of the least squares estimates. We see that the true `limit` coefficient is almost certainly between 0.15 and 0.20.

In contrast, the right hand panel of figure 31 displays contour plots of the RSS associated with possible coefficient estimates for the regression of `balance` onto `limit` and `rating`, which we know to be highly collinear. Now the contours run along a narrow valley; there is a broad range of values for the coefficient estimates that result in equal values for RSS.

Table 11 compares the coefficient estimates obtained from two separate multiple regression models. The first is a regression of `balance` on `age` and `limit`. The second is a regression of `balance` on `rating` and `limit`. In the first regression, both `age` and `limit` are highly significant with very small p-values. In the second, the collinearity between `limit` and `rating` has caused the standard error for the `limit` coefficient to increase by a factor of 12 and the p-value to increase to 0.701. In other words, the importance of the `limit` variable has been masked due to the presence of collinearity.

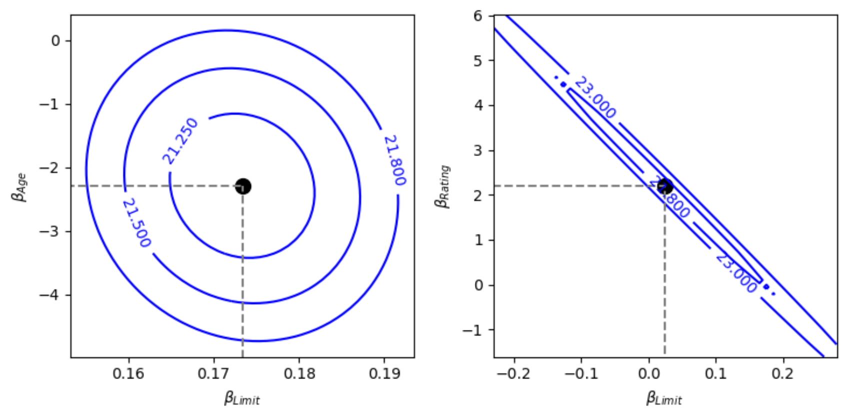

Figure 31: Contour plots for the RSS values as a function of the parameters β for various regressions involving the `Credit` data set. In each plot, the black dots represent the coefficient values corresponding to the minimum RSS. Left: A contour plot of RSS for the regression of `balance` onto `age` and `limit`. The minimum value is well defined. Right: A contour plot of RSS for the regression of `balance` onto `rating` and `limit`. Because of the collinearity, there are many pairs $(\beta_{Limit}, \beta_{Rating})$ with a similar value for RSS.

	Coef.	Std.Err.	t	$P >\| t \|$
Intercept	-173.411	43.828	-3.957	0.0
Age	-2.291	0.672	-3.407	0.001
Limit	0.173	0.005	34.496	0.0
Intercept	-377.537	45.254	-8.343	0.0
Rating	2.202	0.952	2.312	0.021
Limit	0.025	0.064	0.384	0.701

Table 11: The results for two multiple regression models involving the `Credit` data set. The top panel is a regression of `balance` on `age` and `limit`. The bottom panel is a regression of `balance` on `rating` and `limit`. The standard error of $\hat{\beta}_{Limit}$ increases 12-fold in the second regression, due to collinearity.

3.4 The Marketing Plan

3.5 Comparison of Linear Regression with K-Nearest Neighbors

Figure 32 illustrates two KNN fits on a data set with $p = 2$ predictors. The fit with $K = 1$ is shown in the left-hand panel, while the right-hand panel displays the fit with $K = 9$. When $K = 1$, the KNN fit perfectly interpolates the training observations, and consequently takes the form of a step function. When $K = 9$, the KNN fit is still a step function, but averaging over nine observations results in much smaller regions of constant prediction, and consequently a smoother fit.

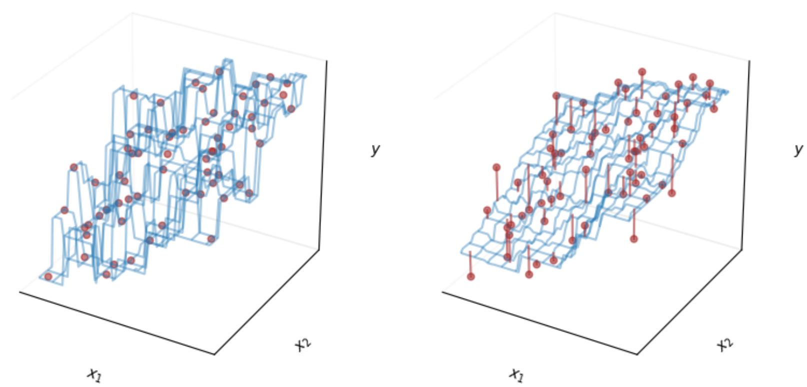

Figure 32: Plots of $\hat{f}(X)$ using KNN regression on two-dimensional data set with 64 observations (brown dots). Left: $K = 1$ results in a rough step function fit. Right: $K = 9$ produces a much smoother fit.

Figure 33 provides an example of KNN regression with data generated from a one-dimensional regression model. the black dashed lines represent $f(X)$, while the blue curves correspond to the KNN fits using $K = 1$ and $K = 9$. In this case, the $K = 1$ predictions are far too variable, while the smoother $K = 9$ fit is much closer to $f(X)$.

Figure 34 represents the linear regression fit to the same data. It is almost perfect. The right hand panel of figure 34 reveals that linear regression outperforms KNN for this data. The green line, plotted as a function of $\frac{1}{K}$, represents the test set mean squared error (MSE) for KNN. The KNN errors are well above the horizontal dashed line, which is the test MSE for linear regression.

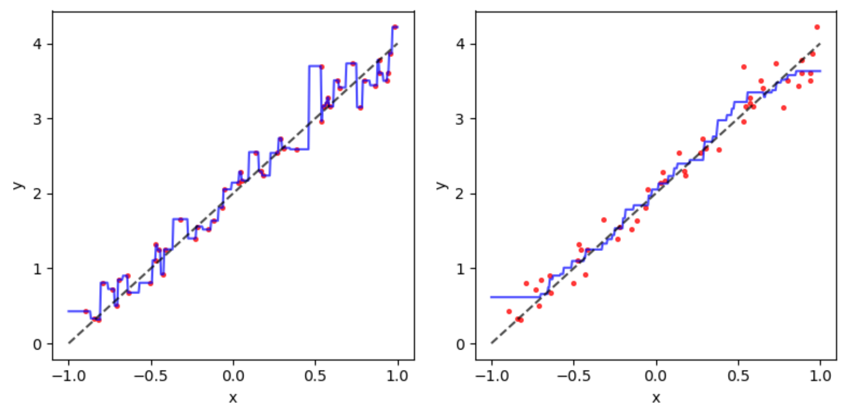

Figure 33: Plots of $\hat{f}(X)$ using KNN regression on a one-dimensional data set with 50 observations. The true relationship is given by the black dashed line. Left: The blue curve corresponds to $K = 1$ and interpolates (i.e., passes directly through) training data. Right: The blue curve corresponds to $K = 9$, and represents a smoother fit.

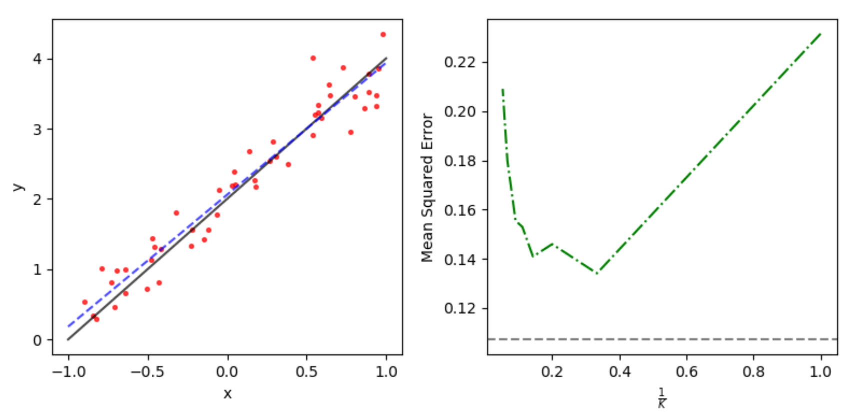

Figure 34: The same data set shown in figure 33 is investigated further. Left: The blue dashed line is the least squares fit to the data. Since $f(X)$ is in fact linear (displayed in black line), the least squares regression line provides a very good estimate of $f(X)$. Right: The dashed horizontal line represents the least squares test set MSE, while the green line corresponds to the MSE for KNN as a function of $\frac{1}{K}$. Linear regression achieves a lower test MSE than does KNN regression, since $f(X)$ is in fact linear.

Figure 35 examines the relative performances of least squares regression and KNN under increasing levels of non-linearity in the relationship between X and Y. In the top row, the true relationship is nearly linear. In this case, we see that the test MSE for linear regression is still superior to that of KNN for low values of K (far right). However, as K increases, KNN outperforms linear regression. The second row illustrates a more substantial deviation from linearity. In this situation, KNN substantially outperforms linear regression for all values of K.

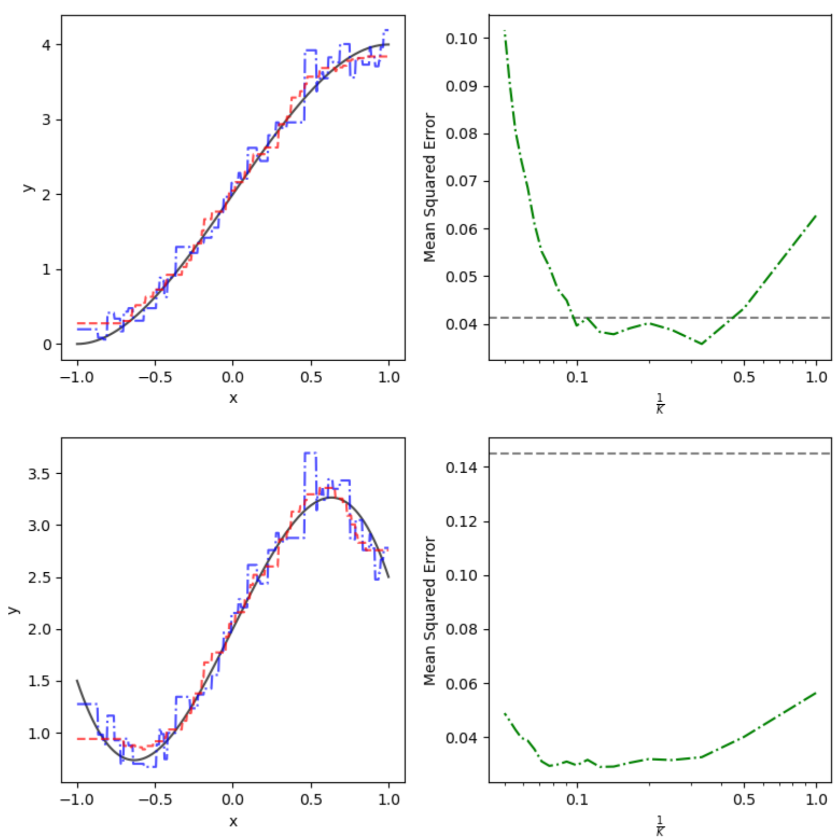

Figure 35: Top Left: In a setting with a slightly non-linear relationship between X and Y (solid black line), the KNN fits with $K = 1$ (blue) and $K = 9$ (red) are displayed. Top Right: For the slightly non-linear data,the test set MSE for least squares regression (horizontal) and KNN with various values of $\frac{1}{K}$ (green) are displayed. Bottom Left and Bottom Right: As in the top panel, but with a strongly non-linear relationship between X and Y.

Figure 36 considers the same strongly non-linear situation as in the lower panel of figure 35, except that we have added additional *noise* predictors that are not associated with the response. When $p = 1$ or $p = 2$, KNN outperforms linear regression. But as we increase p, linear regression becomes superior to KNN. In fact, increase in dimensionality has only caused a small increase in linear regression test set MSE, but it has caused a much bigger increase in the MSE for KNN.

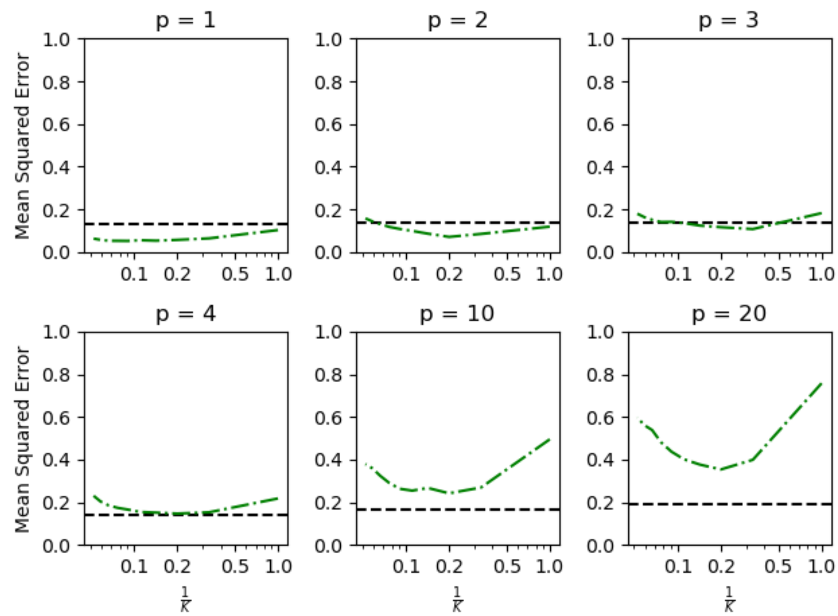

Figure 36: Test MSE for linear regressions (black horizontal lines) and KNN (green curves) as the number of variables p increases. The true function is non-linear in the first variable, as in the lower panel in figure 35, and does not depend upon the additional variables. The performance of linear regression deteriorates slowly in the presense of these additional variables, whereas KNN's performance degrades more quickly as p increases.

3.6 Lab: Linear Regression

3.6.1 Libraries

The `import` function, along with an optional `as`, is used to load *libraries*. Before a library can be loaded, it must be installed on the system.

```
import numpy as np
import statsmodels.formula.api as smf
```

3.6.2 Simple Linear Regression

We load `Boston` data set from `R` library `MASS`. Then we use `ols` function from `statsmodels.formula.api` to fit simple linear regression model, with `medv` as response and `lstat` as the predictor.

Function `summary2()` gives some basic information about the model. We can use `dir()` to find out what other pieces of information are stored in `lm_fit`. The `predict()` function can be used to produce prediction of `medv` for a given value of `lstat`.

```
import statsmodels.formula.api as smf
from statsmodels import datasets

boston = datasets.get_rdataset('Boston', 'MASS').data
print(boston.columns)
print('--------')

lm_reg = smf.ols(formula='medv ~ lstat', data=boston)
lm_fit = lm_reg.fit()
print(lm_fit.summary2())
print('------')

print(dir(lm_fit))
print('------')

print(lm_fit.predict(exog=dict(lstat=[5, 10, 15])))
```

```
Index(['crim', 'zn', 'indus', 'chas', 'nox', 'rm', 'age', 'dis', 'rad', 'tax',
       'ptratio', 'black', 'lstat', 'medv'],
      dtype='object')
--------
                 Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.543
Dependent Variable: medv             AIC:                3286.9750
Date:               2019-05-28 14:10 BIC:                3295.4280
No. Observations:   506              Log-Likelihood:     -1641.5
```

```
Df Model:           1                 F-statistic:        601.6
Df Residuals:       504               Prob (F-statistic): 5.08e-88
R-squared:          0.544             Scale:              38.636
--------------------------------------------------------------------
               Coef.   Std.Err.     t      P>|t|    [0.025   0.975]
--------------------------------------------------------------------
Intercept     34.5538    0.5626   61.4151  0.0000  33.4485  35.6592
lstat         -0.9500    0.0387  -24.5279  0.0000  -1.0261  -0.8740
-------------------------------------------------------------------
Omnibus:             137.043       Durbin-Watson:          0.892
Prob(Omnibus):       0.000         Jarque-Bera (JB):       291.373
Skew:                1.453         Prob(JB):               0.000
Kurtosis:            5.319         Condition No.:          30
===================================================================

------
['HC0_se', 'HC1_se', 'HC2_se', 'HC3_se', '_HCCM', '__class__', '__delattr__',
'__dict__', '__dir__', '__doc__', '__eq__', '__format__', '__ge__',
'__getattribute__', '__gt__', '__hash__', '__init__', '__init_subclass__',
'__le__', '__lt__', '__module__', '__ne__', '__new__', '__reduce__',
'__reduce_ex__', '__repr__', '__setattr__', '__sizeof__', '__str__',
'__subclasshook__', '__weakref__', '_cache', '_data_attr',
'_get_robustcov_results', '_is_nested', '_wexog_singular_values', 'aic',
'bic', 'bse', 'centered_tss', 'compare_f_test', 'compare_lm_test',
'compare_lr_test', 'condition_number', 'conf_int', 'conf_int_el', 'cov_HC0',
'cov_HC1', 'cov_HC2', 'cov_HC3', 'cov_kwds', 'cov_params', 'cov_type',
'df_model', 'df_resid', 'eigenvals', 'el_test', 'ess', 'f_pvalue', 'f_test',
'fittedvalues', 'fvalue', 'get_influence', 'get_prediction',
'get_robustcov_results', 'initialize', 'k_constant', 'llf', 'load', 'model',
'mse_model', 'mse_resid', 'mse_total', 'nobs', 'normalized_cov_params',
'outlier_test', 'params', 'predict', 'pvalues', 'remove_data', 'resid',
'resid_pearson', 'rsquared', 'rsquared_adj', 'save', 'scale', 'ssr',
'summary', 'summary2', 't_test', 't_test_pairwise', 'tvalues',
'uncentered_tss', 'use_t', 'wald_test', 'wald_test_terms', 'wresid']
------
0    29.803594
1    25.053347
2    20.303101
dtype: float64
```

We will now plot `medv` and `lstat` along with least squares regression line.

```
import statsmodels.formula.api as smf
from statsmodels import datasets

boston = datasets.get_rdataset('Boston', 'MASS').data
print(boston.columns)
print('--------')

lm_reg = smf.ols(formula='medv ~ lstat', data=boston)
lm_fit = lm_reg.fit()
print(lm_fit.summary2())
print('------')

print(dir(lm_fit))
print('------')

print(lm_fit.predict(exog=dict(lstat=[5, 10, 15])))
import statsmodels.api as sm
import matplotlib.pyplot as plt

fig = plt.figure()
ax = fig.add_subplot(111)
boston.plot(x='lstat', y='medv', alpha=0.7, ax=ax)
sm.graphics.abline_plot(model_results=lm_fit, ax=ax, c='r')
```

Next we examine some diagnostic plots.

```
import statsmodels.formula.api as smf
from statsmodels import datasets

boston = datasets.get_rdataset('Boston', 'MASS').data
print(boston.columns)
print('--------')

lm_reg = smf.ols(formula='medv ~ lstat', data=boston)
lm_fit = lm_reg.fit()
print(lm_fit.summary2())
print('------')
```

```
print(dir(lm_fit))
print('------')

print(lm_fit.predict(exog=dict(lstat=[5, 10, 15])))
import statsmodels.api as sm
from statsmodels.nonparametric.smoothers_lowess import lowess
import matplotlib.pyplot as plt
import numpy as np

fig = plt.figure()
ax1 = fig.add_subplot(221)
ax1.scatter(lm_fit.fittedvalues, lm_fit.resid, s=5, c='b', alpha=0.6)
ax1.axhline(y=0, linestyle='--', c='r')
# resid_lowess_fit = lowess(endog=lm_fit.resid, exog=lm_fit.fittedvalues,
#                            is_sorted=True)
# ax1.plot(resid_lowess_fit[:,0], resid_lowess_fit[:,1])
ax1.set_xlabel('Fitted values')
ax1.set_ylabel('Residuals')
ax1.set_title('Residuals vs Fitted')

ax2=fig.add_subplot(222)
sm.graphics.qqplot(lm_fit.resid, ax=ax2, markersize=3, line='s',
   linestyle='--', fit=True, alpha=0.4)
ax2.set_ylabel('Standardized residuals')
ax2.set_title('Normal Q-Q')

influence = lm_fit.get_influence()
standardized_resid = influence.resid_studentized_internal
ax3 = fig.add_subplot(223)
ax3.scatter(lm_fit.fittedvalues, np.sqrt(np.abs(standardized_resid)), s=5,
    alpha=0.4, c='b')
ax3.set_xlabel('Fitted values')
ax3.set_ylabel(r'$\sqrt{\mid Standardized\; residuals \mid}$')
ax3.set_title('Scale-Location')

ax4 = fig.add_subplot(224)
sm.graphics.influence_plot(lm_fit, size=2, alpha=0.4, c='b',  ax=ax4)
ax4.xaxis.label.set_size(10)
ax4.yaxis.label.set_size(10)
```

```
ax4.title.set_size(12)
ax4.set_xlim(0, 0.03)
for txt in ax4.texts:
    txt.set_visible(False)
ax4.axhline(y=0, linestyle='--', color='grey')

fig.tight_layout()
```

3.6.3 Multiple Linear Regression

In order to fit a multiple regression model using least squares, we again use the `ols` and `fit` functions. The syntax `ols(formula='y ~ x1 + x2 + x3')` is used to fit a model with three predictors, `x1`, `x2`, and `x3`. The `summary2()` now outputs the regression coefficients for all three predictors.

`statsmodels` does not seem to have `R` like facility to include all variables using the formula `y ~ .`. To include all variables, we either write them individually, or use code to create a formula.

```
import statsmodels.formula.api as smf
from statsmodels import datasets

boston = datasets.get_rdataset('Boston', 'MASS').data

lm_reg = smf.ols(formula='medv ~ lstat + age', data=boston)
lm_fit = lm_reg.fit()

print(lm_fit.summary2())
print('--------')

# Create formula to include all variables
all_columns = list(boston.columns)
all_columns.remove('medv')
my_formula = 'medv ~ ' + ' + '.join(all_columns)
print(my_formula)
print('--------')

all_reg = smf.ols(formula=my_formula, data=boston)
all_fit = all_reg.fit()
print(all_fit.summary2())
print('--------')
```

```
                 Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.549
Dependent Variable: medv             AIC:                3281.0064
Date:               2019-05-29 10:07 BIC:                3293.6860
No. Observations:   506              Log-Likelihood:     -1637.5
Df Model:           2                F-statistic:        309.0
Df Residuals:       503              Prob (F-statistic): 2.98e-88
R-squared:          0.551            Scale:              38.108
-------------------------------------------------------------------
               Coef.   Std.Err.     t      P>|t|    [0.025   0.975]
-------------------------------------------------------------------
Intercept     33.2228    0.7308   45.4579  0.0000  31.7869  34.6586
lstat         -1.0321    0.0482  -21.4163  0.0000  -1.1267  -0.9374
age            0.0345    0.0122    2.8256  0.0049   0.0105   0.0586
------------------------------------------------------------------
Omnibus:             124.288       Durbin-Watson:          0.945
Prob(Omnibus):       0.000         Jarque-Bera (JB):       244.026
Skew:                1.362         Prob(JB):               0.000
Kurtosis:            5.038         Condition No.:          201
==================================================================


--------
medv ~ crim + zn + indus + chas + nox + rm + age + dis + rad + tax +
ptratio + black + lstat
--------
                 Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.734
Dependent Variable: medv             AIC:                3025.6086
Date:               2019-05-29 10:07 BIC:                3084.7801
No. Observations:   506              Log-Likelihood:     -1498.8
Df Model:           13               F-statistic:        108.1
Df Residuals:       492              Prob (F-statistic): 6.72e-135
R-squared:          0.741            Scale:              22.518
-------------------------------------------------------------------
            Coef.   Std.Err.     t      P>|t|    [0.025    0.975]
-------------------------------------------------------------------
Intercept  36.4595    5.1035    7.1441  0.0000   26.4322   46.4868
crim       -0.1080    0.0329   -3.2865  0.0011   -0.1726   -0.0434
```

```
zn           0.0464    0.0137    3.3816  0.0008    0.0194    0.0734
indus        0.0206    0.0615    0.3343  0.7383   -0.1003    0.1414
chas         2.6867    0.8616    3.1184  0.0019    0.9939    4.3796
nox        -17.7666    3.8197   -4.6513  0.0000  -25.2716  -10.2616
rm           3.8099    0.4179    9.1161  0.0000    2.9887    4.6310
age          0.0007    0.0132    0.0524  0.9582   -0.0253    0.0266
dis         -1.4756    0.1995   -7.3980  0.0000   -1.8675   -1.0837
rad          0.3060    0.0663    4.6129  0.0000    0.1757    0.4364
tax         -0.0123    0.0038   -3.2800  0.0011   -0.0197   -0.0049
ptratio     -0.9527    0.1308   -7.2825  0.0000   -1.2098   -0.6957
black        0.0093    0.0027    3.4668  0.0006    0.0040    0.0146
lstat       -0.5248    0.0507  -10.3471  0.0000   -0.6244   -0.4251
-----------------------------------------------------------------
Omnibus:             178.041       Durbin-Watson:          1.078
Prob(Omnibus):       0.000         Jarque-Bera (JB):       783.126
Skew:                1.521         Prob(JB):               0.000
Kurtosis:            8.281         Condition No.:          15114
=================================================================
* The condition number is large (2e+04). This might indicate
strong multicollinearity or other numerical problems.
--------
```

3.6.4 Interaction Terms

The syntax `lstat:black` tells `ols` to include an interaction term between `lstat` and `black`. The syntax `lstat*age` simultaneously includes `lstat`, `age`, and the interaction term lstat×

```
import statsmodels.formula.api as smf
from statsmodels import datasets

boston = datasets.get_rdataset('Boston', 'MASS').data

my_reg = smf.ols(formula='medv ~ lstat * age', data=boston)
my_fit = my_reg.fit()
print(my_fit.summary2())
```

```
                 Results: Ordinary least squares
=================================================================
Model:              OLS              Adj. R-squared:     0.553
Dependent Variable: medv             AIC:                3277.9547
```

```
Date:               2019-05-29 11:48 BIC:                 3294.8609
No. Observations:   506              Log-Likelihood:      -1635.0
Df Model:           3                F-statistic:         209.3
Df Residuals:       502              Prob (F-statistic): 4.86e-88
R-squared:          0.556            Scale:               37.804
-------------------------------------------------------------------
               Coef.  Std.Err.    t     P>|t|    [0.025   0.975]
-------------------------------------------------------------------
Intercept     36.0885    1.4698 24.5528 0.0000 33.2007  38.9763
lstat         -1.3921    0.1675 -8.3134 0.0000 -1.7211  -1.0631
age           -0.0007    0.0199 -0.0363 0.9711 -0.0398   0.0383
lstat:age      0.0042    0.0019  2.2443 0.0252  0.0005   0.0078
------------------------------------------------------------------
Omnibus:              135.601      Durbin-Watson:          0.965
Prob(Omnibus):        0.000        Jarque-Bera (JB):       296.955
Skew:                 1.417        Prob(JB):               0.000
Kurtosis:             5.461        Condition No.:          6878
==================================================================
* The condition number is large (7e+03). This might indicate
strong multicollinearity or other numerical problems.
```

3.6.5 Non-linear Transformations of the Predictors

The `ols` function can also accommodate non-linear transformations of the predictors. For example, given a predictor X, we can create predictor X^2 using `I(X ** 2)`. We now perform a regression of `medv` onto `lstat` and `lstat`^{2}.

The near-zero p-value associated with the quadratic term suggests that it leads to an improve model. We use `anova_lm()` function to further quantify the extent to which the quadratic fit is superior to the linear fit. The null hypothesis is that the two models fit the data equally well. The alternative hypothesis is that the full model is superior. Given the large F-statistic and zero p-value, this provides very clear evidence that the model with quadratic term is superior. A plot of residuals versus fitted values shows that, with quadratic term included, there is no discernible pattern in residuals.

```
import statsmodels.formula.api as smf
from statsmodels import datasets
import statsmodels.api as sm
lowess = sm.nonparametric.lowess
```

```
import matplotlib.pyplot as plt

boston = datasets.get_rdataset('Boston', 'MASS').data

my_reg = smf.ols(formula='medv ~ lstat', data=boston)
my_fit = my_reg.fit()

my_reg2 = smf.ols(formula='medv ~ lstat + I(lstat ** 2)', data=boston)
my_fit2 = my_reg2.fit()
print(my_fit.summary2())
print('--------')

print(sm.stats.anova_lm(my_fit2))
print('--------')

print(sm.stats.anova_lm(my_fit, my_fit2))

my_regs = (my_reg, my_reg2)

fig = plt.figure(figsize=(8,4))
i_reg = 1
for reg in my_regs:
    ax = fig.add_subplot(1, 2, i_reg)
    fit = reg.fit()
    ax.scatter(fit.fittedvalues, fit.resid, s=7, alpha=0.6)
    lowess_fit = lowess(fit.resid, fit.fittedvalues)
    ax.plot(lowess_fit[:,0], lowess_fit[:,1], c='r')
    ax.axhline(y=0, linestyle='--', color='grey')
    ax.set_xlabel('Fitted values')
    ax.set_ylabel('Residuals')
    ax.set_title(reg.formula)
    i_reg += 1

fig.tight_layout()
```

```
                 Results: Ordinary least squares
==================================================================
Model:              OLS              Adj. R-squared:     0.543
Dependent Variable: medv             AIC:                3286.9750
Date:               2019-05-29 12:41 BIC:                3295.4280
```

```
No. Observations:   506             Log-Likelihood:     -1641.5
Df Model:           1               F-statistic:        601.6
Df Residuals:       504             Prob (F-statistic): 5.08e-88
R-squared:          0.544           Scale:              38.636
-------------------------------------------------------------------
              Coef.   Std.Err.     t      P>|t|    [0.025   0.975]
-------------------------------------------------------------------
Intercept    34.5538    0.5626   61.4151  0.0000  33.4485  35.6592
lstat        -0.9500    0.0387  -24.5279  0.0000  -1.0261  -0.8740
------------------------------------------------------------------
Omnibus:              137.043      Durbin-Watson:          0.892
Prob(Omnibus):        0.000        Jarque-Bera (JB):       291.373
Skew:                 1.453        Prob(JB):               0.000
Kurtosis:             5.319        Condition No.:          30
==================================================================

--------
                 df        sum_sq       mean_sq           F         PR(>F)
lstat           1.0  23243.913997  23243.913997  761.810354  8.819026e-103
I(lstat ** 2)   1.0   4125.138260   4125.138260  135.199822   7.630116e-28
Residual      503.0  15347.243158     30.511418         NaN            NaN
--------
   df_resid           ssr  df_diff     ss_diff           F        Pr(>F)
0     504.0  19472.381418      0.0         NaN         NaN           NaN
1     503.0  15347.243158      1.0  4125.13826  135.199822  7.630116e-28
```

3.6.6 Qualitative Predictors

We will now examine `Carseats` data, which is part of the `ISLR` library. We will attempt to predict `Sales` (child car seat sales) based on a number of predictors. `statsmodels` automatically converts string variables into categorical variables. If we want `statsmodels` to treat a numerical variable `x` as qualitative predictor, the formula should be `y ~ C(x)`. Here `C()` stands for categorical.

```
import statsmodels.formula.api as smf
from statsmodels import datasets

carseats = datasets.get_rdataset('Carseats', 'ISLR').data
print(carseats.columns)
```

```
print('--------')

all_columns = list(carseats.columns)
all_columns.remove('Sales')
my_formula = 'Sales ~ ' + ' + '.join(all_columns)
my_formula +=  ' + Income:Advertising + Price:Age'

print(my_formula)
print('--------')

my_reg = smf.ols(formula=my_formula, data=carseats)
my_fit = my_reg.fit()
print(my_fit.summary2())
```

```
Index(['Sales', 'CompPrice', 'Income', 'Advertising', 'Population', 'Price',
       'ShelveLoc', 'Age', 'Education', 'Urban', 'US'],
      dtype='object')
--------
Sales ~ CompPrice + Income + Advertising + Population + Price + ShelveLoc + Age + Educa
--------
                  Results: Ordinary least squares
====================================================================
Model:              OLS              Adj. R-squared:     0.872
Dependent Variable: Sales            AIC:                1157.3378
Date:               2019-05-29 12:53 BIC:                1213.2183
No. Observations:   400              Log-Likelihood:     -564.67
Df Model:           13               F-statistic:        210.0
Df Residuals:       386              Prob (F-statistic): 6.14e-166
R-squared:          0.876            Scale:              1.0213
--------------------------------------------------------------------
                     Coef.  Std.Err.    t     P>|t|   [0.025  0.975]
--------------------------------------------------------------------
Intercept            6.5756   1.0087   6.5185 0.0000  4.5922  8.5589
ShelveLoc[T.Good]    4.8487   0.1528  31.7243 0.0000  4.5482  5.1492
ShelveLoc[T.Medium]  1.9533   0.1258  15.5307 0.0000  1.7060  2.2005
Urban[T.Yes]         0.1402   0.1124   1.2470 0.2132 -0.0808  0.3612
US[T.Yes]           -0.1576   0.1489  -1.0580 0.2907 -0.4504  0.1352
CompPrice            0.0929   0.0041  22.5668 0.0000  0.0848  0.1010
Income               0.0109   0.0026   4.1828 0.0000  0.0058  0.0160
Advertising          0.0702   0.0226   3.1070 0.0020  0.0258  0.1147
```

```
Population              0.0002   0.0004   0.4329 0.6653 -0.0006  0.0009
Price                  -0.1008   0.0074 -13.5494 0.0000 -0.1154 -0.0862
Age                    -0.0579   0.0160  -3.6329 0.0003 -0.0893 -0.0266
Education              -0.0209   0.0196  -1.0632 0.2884 -0.0594  0.0177
Income:Advertising      0.0008   0.0003   2.6976 0.0073  0.0002  0.0013
Price:Age               0.0001   0.0001   0.8007 0.4238 -0.0002  0.0004
---------------------------------------------------------------------
Omnibus:                   1.281        Durbin-Watson:           2.047
Prob(Omnibus):             0.527        Jarque-Bera (JB):        1.147
Skew:                      0.129        Prob(JB):                0.564
Kurtosis:                  3.050        Condition No.:          130576
=====================================================================
* The condition number is large (1e+05). This might indicate
strong multicollinearity or other numerical problems.
```

3.6.7 Calling `R` from `Python`

4 Classification

4.1 An Overview of Classification

In figure 37, we have plotted annual `income` and monthly credit card `balance` for a subset of individuals in `Credit` data set. The left hand panel displays individuals who defaulted in brown, and those who did not in blue. We have plotted only a fraction of individuals who did not default. It appears that individuals who defaulted tended to have higher credit card balances than those who did not. In the right hand panel, we show two pairs of boxplots. The first shows the distribution of `balance` split by the binary `default` variable; the second is a similar plot for `income`.

4.2 Why Not Linear Regression?

4.3 Logistic Regression

Using `Default` data set, in figure 38 we show probability of default as a function of `balance`. The left panel shows a model fitted using linear regression. Some of the probabilities estimates (for low balance) are outside the $[0, 1]$ interval. The right panel shows a model fitted using logistic regression, which models the probability of default as a function of `balance`. Now all probability estimates are in the $[0, 1]$ interval.

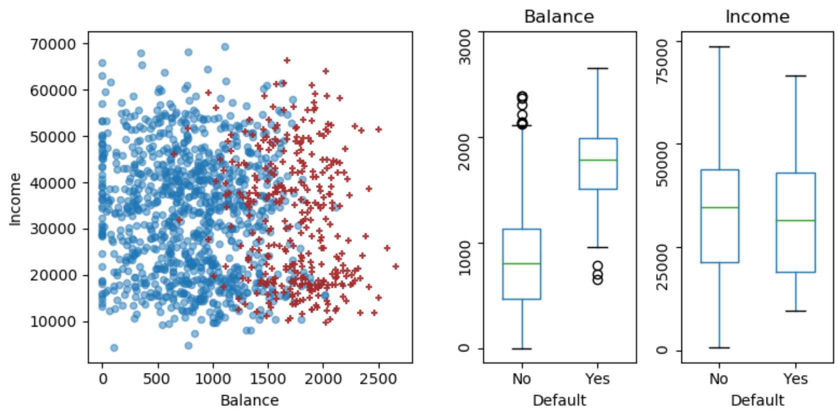

Figure 37: The `Default` data set. Left: The annual income and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card debt are shown in brown, and those who did not default are shown in blue. Center: Boxplots of `balance` as a function of `default` status. Right: Boxplots of `income` as a function of `default` status.

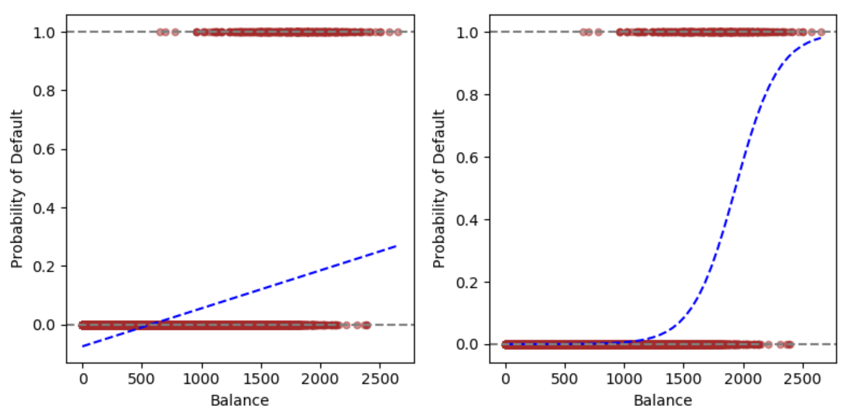

Figure 38: Classification using `Default` data. Left: Estimated probability of `default` using linear regression. Some estimated probabilities are negative! The brown ticks indicate the 0/1 values coded for `default` (`No` or `Yes`). Right: Predicted probabilities of `default` using logistic regression. All probabilities lie between 0 and 1.

Table 12 shows the coefficient estimates and related information that result from fitting a logistic regression model on the `Default` data in order to predict the probability of `default = Yes` using `balance`.

	Coef.	Std.Err.	z	$P >\| z \|$
Intercept	-10.6513	0.3612	-29.4913	0.0
balance	0.0055	0.0002	24.9524	0.0

Table 12: For the `Default` data, estimated coefficients of the logistic regression model that predicts the probability of `default` using `balance`. A one-unit increase in `balance` is associated with an increase in the log odds of `default` by 0.0055 units.

Table 13 shows the results of logistic model where `default` is a function of the qualitative variable `student`.

Table 14 shows the coefficient estimates for a logistic regression model that uses `balance`, `income` (in thousands of dollars), and `student` status to predict probability of `default`.

	Coef.	Std.Err.	z	$P >\| z \|$
Intercept	-3.5041	0.0707	-49.5541	0.0
student[T.Yes]	0.4049	0.115	3.5202	0.0004

Table 13: For the `Default` data, estimated coefficients of the logistic regression model that predicts the probability of `default` using student status.

	Coef.	Std.Err.	z	$P >\| z \|$
Intercept	-10.869	0.4923	-22.0793	0.0
student[T.Yes]	-0.6468	0.2363	-2.7376	0.0062
balance	0.0057	0.0002	24.7365	0.0
income	0.003	0.0082	0.3698	0.7115

Table 14: For the `Default` data, estimated coefficients of the logistic regression model that predicts the probability of `default` using `balance`, `income`, and `student` status. In fitting this model, `income` was measured in thousands of dollars.

The left hand panel of figure 39 shows average default rates for students and non-students, respectively, as a function of credit card balance. *For a fixed value* of `balance` and `income`, a student is less likely to default than a non-student. This is true for all values of balance. This is consistent with negative coefficient of student in table 14. But the horizontal lines near the base of the plot, which show the default rates for students and non-students averaged over all values of `balance` and `income`, suggest the opposite effect: the overall student default rate is higher than non-student default rate. Consequently, there is a positive coefficient for `student` in the single variable logistic regression output shown in table 13.

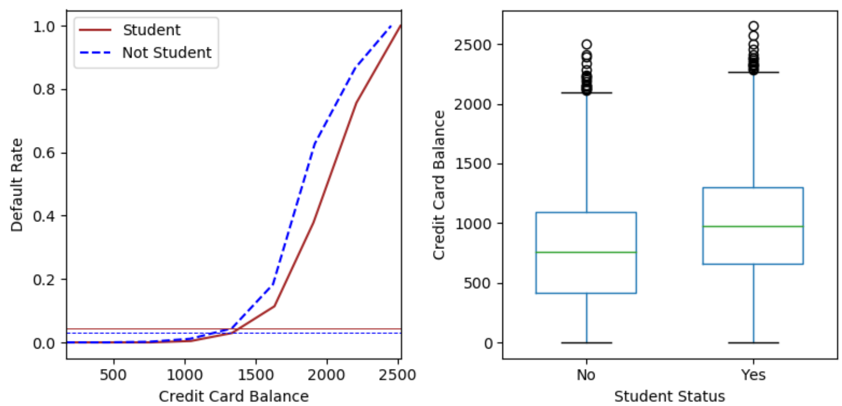

Figure 39: Confounding in the `Default` data. Left: Default rates are shown for students (brown) and non-students (blue). The solid lines display default rate as a function of `balance`, while the horizontal lines display the overall default rates. Right: Boxplots of `balance` for students and non-students are shown.

4.4 Linear Discriminant Analysis

In the left panel of figure 40, two normal density functions that are displayed, $f_1(x)$ and $f_2(x)$, represent two distinct classes. The Bayes classifier boundary, shown as vertical dashed line, is estimated using the function `GaussianNB()`. The right hand panel displays a histogram of a random sample of 20 observations from each class. The LDA decision boundary is shown as firm vertical line.

Two examples of multivariate Gaussian distributions with $p = 2$ are shown in figure 41. In the upper panel, the height of the surface at any particular point represents the probability that both X_1 and X_2 fall in the small region around that point. If the surface is cut along the X_1 axis or along the X_2 axis, the resulting cross-section will have the shape of a one-dimensional normal distribution. The left-hand panel illustrates an example in which

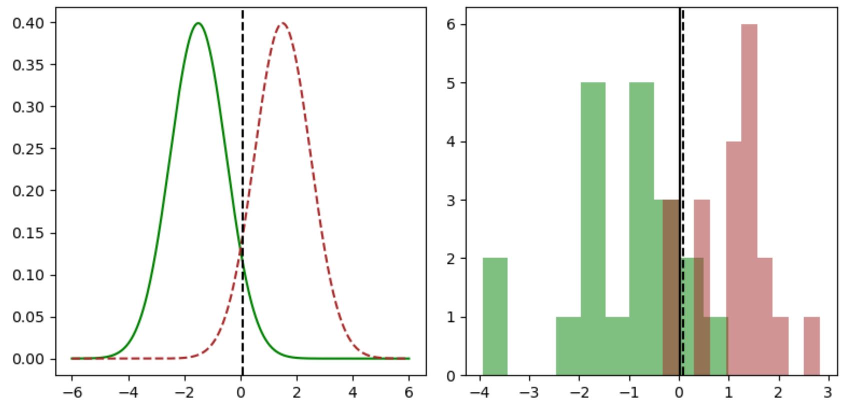

Figure 40: Left: Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary. Right: 20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.

$\text{var}(X_1) = \text{var}(X_2)$ and $\text{cor}(X_1, X_2) = 0$; this surface has a characteristic *bell shape*. However, the bell shape will be distorted if the predictors are correlated or have unequal variances, as is illustrated in the right-hand panel of figure 41. In this situation, the base of the bell will have an elliptical, rather than circular, shape. The contour plots in the lower panel are not in the book.

Figure 42 shows an example of three equally sized Gaussian classes with class-specific mean vectors and a common covariance matrix. The dashed lines are the Bayes decision boundaries.

A *confusion matrix*, shown for the `Default` data in table 15, is a convenient way to display prediction of default in comparison to true default. Table 16 shows the error rates that result when we label any customer with a posterior probability of default above 20% to the *default* class.

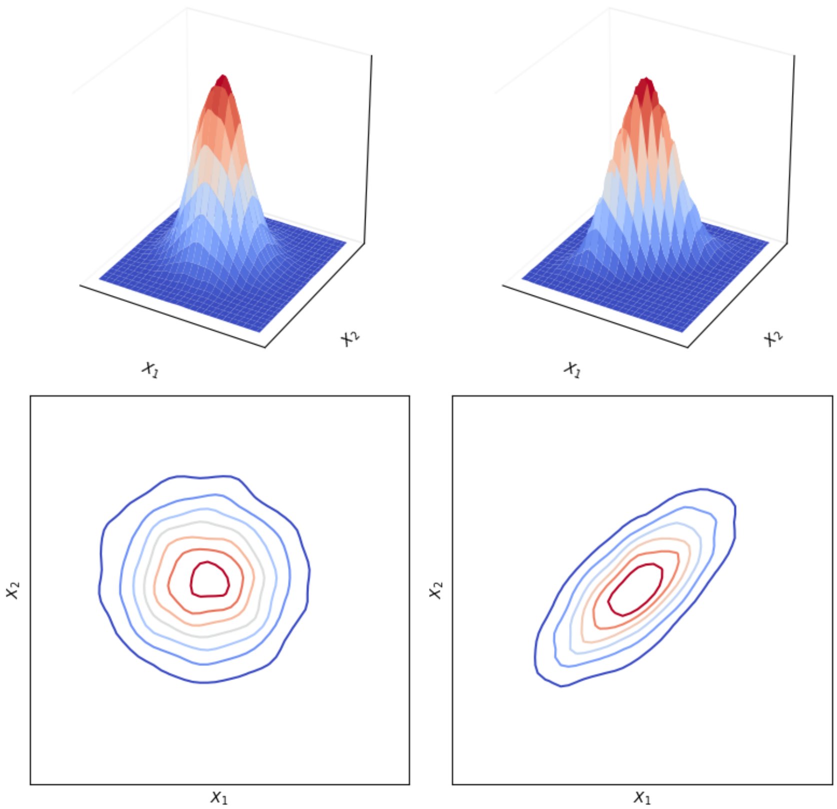

Figure 41: Two multivariate Gaussian density functions are shown, with $p = 2$. Left: The two predictors are uncorrelated. Right: The two predictors have a correlation of 0.7. The lower panel shows contour plots of the surfaces drawn in the upper panel. Here the correlations can be easily seen.

	true No	true Yes	Total
predict No	9645	254	9899
predict Yes	22	79	101
Total	9667	333	10000

Table 15: A confusion matrix compares the LDA predictions to the true default statuses for the training observations in the `Default` data set. Elements of the diagonal matrix represent individuals whose default statuses were correctly predicted, while off-diagonal elements represent individuals that were missclassified.

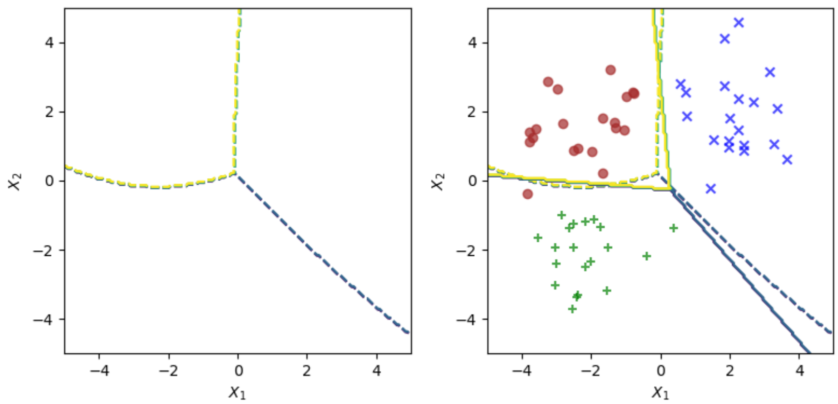

Figure 42: An example with three classes. The observation from each class are drawn from a multivariate Gaussian distribution with $p = 2$, with a class-specific mean vector and a common covariance matrix. Left: The dashed lines are the Bayes decision boundaries. Right: 20 observations were generated from each class, and the corresponding LDA decision boundaries are indicated using solid black lines. The Bayes decision boundaries are once again shown as dashed lines.

	true No	true Yes	Total
predict No	9435	140	9575
predict Yes	232	193	425
Total	9667	333	10000

Table 16: A confusion matrix compares LDA predictions to the true default statuses for the training observations in the `Default` data set, using a modified threshold value that predicts default for any individuals whose posterior default probability exceeds 20%.