Introduction to Statistical Learning
with applications in Python

Based on "Introduction to Statistical Learning, with applications in R" by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

Linear Regression, Part 3

qualitative predictors, interaction & non-linear extensions, outliers, high leverage points

Kurt Rinnert

Physics Without Frontiers

The Abdus Salam International Centre for Theoretical Physics

University of Liverpool

Copyright © 2019
Kurt Rinnert <kurt.rinnert@cern.ch>, Kate Shaw <kshaw@ictp.it>
Copying and distribution of this file, with or without modification, are permitted in any medium without royalty provided the copyright notice and this notice are preserved. This file is offered as-is, without any warranty.
Some of the figures in this presentation 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

Abstract

Linear models are an important topic in statistical learning.
The true relationships between predictors and responses are rarely linear. But linear models often provide reasonable approximation. They provide high interpretability and have low variance, mitigating the risk of over-fitting. Linear models can be extended to include (some) non-linear relationships.
Linear models also provide an excellent baseline to compare other models against: if our sophisticated model does not do much better than a linear model we might consider trading some bias for lower variance.

Overview

- Qualitative versus quantitative predictors.
- Interactions among predictors.
- Non-linear extensions to the linear model.
- Outliers.
- High leverage points.

This will conclude our long journey through linear regression.

Example: Credit Data Set

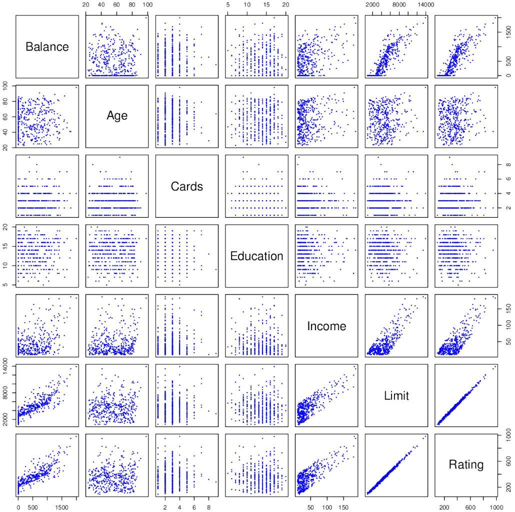
- The plot on the right shows the *quantitative* variables in the data set.
- The dataset also contains *qualitative* predictors:

  `gender`: `Male`, `Female`

  `married`: `Yes`, `No`

  `student`: `Yes`, `No`

  `ethnicity`: `Asian`, `African American`, `Caucasian`



We need to somehow *encode* the qualitative predictors.

Qualitative Predictors with Two Levels

- The variables `gender`, `student` and `married` are qualitative predictors with two levels.
- Suppose we are interested in whether `gender` influences `balance`.
- We can define a *dummy variable* to encode the gender:

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is } \texttt{Female} \\ 0 & \text{if } i\text{th person is } \texttt{Male} \end{cases}$$

- This results in the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is } \texttt{Female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is } \texttt{Male} \end{cases}$$

The distinction is important for the interpretation of the model.

Qualitative Predictors with Two Levels

Model Interpretation

$$\begin{aligned} \beta_0 &: \text{average } \texttt{balance} \text{ among males} \\ \beta_0 + \beta_1 &: \text{average } \texttt{balance} \text{ among females} \\ \beta_1 &: \text{average } \texttt{balance} \text{ difference between females and males} \end{aligned}$$

Fit Result

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

The observed average difference of $19.73 is *not* significant.

Qualitative Predictors with Two Levels

- We can also encode the gender differently:

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is } \texttt{Female} \\ -1 & \text{if } i\text{th person is } \texttt{Male} \end{cases}$$

- This results in the model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is } \texttt{Female} \\ \beta_0 - \beta_1 + \epsilon_i & \text{if } i\text{th person is } \texttt{Male} \end{cases}$$

This leads to a different interpretation of the coefficients.

Qualitative Predictors with Two Levels

Model Interpretation

β_0 : overall average `balance`, disregarding gender

β_1 : amount above (below) average for females (males)

Fit Result

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	519.67	23.03	22.569	< 0.0001
`gender`	9.87	23.03	0.429	0.6690

Note that the fit is essentially the same as before.

Qualitative Predictors with more than Two Levels

- The `ethnicity` variable has three possible values.
- In this case we need two dummy variables for the encoding, x_{i1} and x_{i2}.
- We choose the value `African American` as the *baseline*:

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is } \texttt{Asian} \\ 0 & \text{if } i\text{th person is not } \texttt{Asian} \end{cases} \qquad x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is } \texttt{Caucasian} \\ 0 & \text{if } i\text{th person is not } \texttt{Caucasian} \end{cases}$$

- This results in the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is } \texttt{Asian} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is } \texttt{Caucasian} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is } \texttt{African American} \end{cases}$$

The choice of the baseline is arbitrary.

Qualitative Predictors with more than Two Levels

Model Interpretation

β_0 : average `balance` among African Americans

β_1 : average `balance` difference between African Americans and Asians

β_2 : average `balance` difference between African Americans and Caucasians

Fit Result

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	531.00	46.32	11.464	< 0.0001
`ethnicity[Asian]`	−18.69	65.02	−0.287	0.7740
`ethnicity[Caucasian]`	−12.50	56.68	−0.221	0.8260

We always need one less dummy variable than there are levels.

Extensions of the Linear Model

- The standard linear regression model provides easily interpretable results.
- It also works well on many real world problems (event though the true relationships are rarely linear).
- However, it has the restrictive assumptions of *additivity* and *linearity*:
    1. Additive assumption: The effect of a predictor X_j on the response Y is independent of the other predictors.
    2. Linearity assumption: The change in the response Y due to a one-unit change in X_j is constant.

We explore some ways to weaken these assumptions while keeping the model linear.

Removing the Additive Assumption

- We explore the idea of *interaction terms* using the `Advertising` data set.
- In particular, we want to check whether there is some synergy between the `TV` and `radio` budgets.
- Recall the form of the additive model only including the *main effects*:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon$$

- We now drop the additive assumption and introduce an interaction term:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{TV} \times \texttt{radio} + \epsilon$$

This model is still linear in the parameters $\boldsymbol{\beta}$. Always include the *main effects*.

Removing the Additive Assumption

- We can write the model in a slightly different form.

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \\
&= \beta_0 + (\beta_1 + \beta_3 x_2) x_1 + \beta_2 x_2 + \epsilon \\
&= \beta_0 + \tilde{\beta}_1 x_1 + \beta_2 x_2 + \epsilon
\end{aligned}
$$

$$
\begin{aligned}
\texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{TV} \times \texttt{radio} + \epsilon \\
&= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon \\
&= \beta_0 + \tilde{\beta}_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon
\end{aligned}
$$

- We can now interpret β_3 as the increase of the influence of TV due to radio (or vice versa).

Remember we do not make formal claims of causality.

Removing the Additive Assumption

Fit Result ($R^2_{\text{interaction}} = 0.968,\ R^2_{\text{additive}} = 0.897$)

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

There is strong evidence of synergy between the two budgets.

Interactions with Qualitative Predictors

- Interaction terms with qualitative predictors are also possible.
- They even have a particularly nice interpretation.
- We return to the `Credit` data set to examine this.
- Suppose we want to predict `sales` from `income` (quantitative) and `student` (qualitative):

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

This amounts to fitting two parallel lines: one for students and one for non-students.

Interactions with Qualitative Predictors

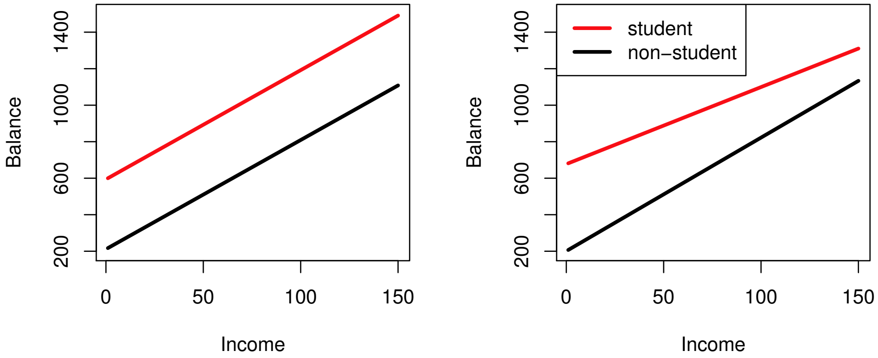
- We want to allow for the effect of `income` to change, depending on whether the value of `student` is `Yes` or `No`.
- We therefore introduce an interaction term and the model becomes:

$$\texttt{balance}_i \approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 & \text{if } i\text{th person is a student} \\ 0 & \text{if } i\text{th person is not a student} \end{cases}$$

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

The lines are no longer parallel: the slope now depends on the value of `student`.

Interactions with Qualitative Predictors



Left: additive model, Right: model with interaction term.

In the model with interaction the slopes are different.

Non-linear Relationships

- We now look into the assumption that the relationship between the predictors X and the response Y is linear.
- We already stressed many times that this is in general *not* true.
- A common approach to address this is to add non-linear functions of one or more predictors to the model design.
- In principle, this can be *any* function, of *any* number of predictors.
- Often we restrict ourselves to *polynomial* functions.
- It is important to note that the model is still linear in the parameters β.

This idea extends beyond linear models and is sometimes called *feature engineering*.

Non-linear Relationships

- We illustrate this on the `Auto` data set.
- We introduce a quadratic term in an attempt to better predict `mpg` from `horsepower`:

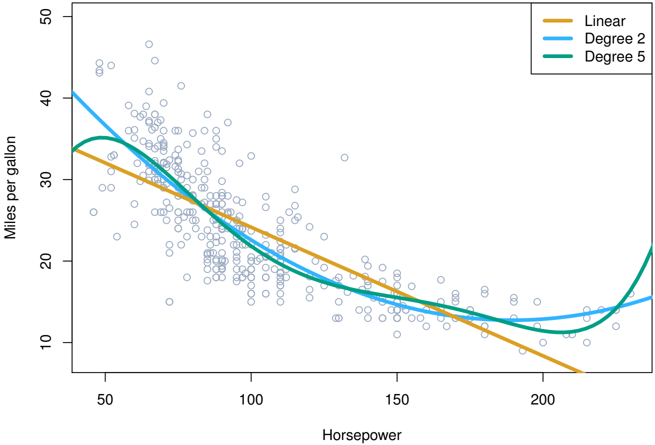
$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

- This results in the following fit:

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	56.9001	1.8004	31.6	< 0.0001
`horsepower`	-0.4662	0.0311	-15.0	< 0.0001
$\texttt{horsepower}^2$	0.0012	0.0001	10.1	< 0.0001

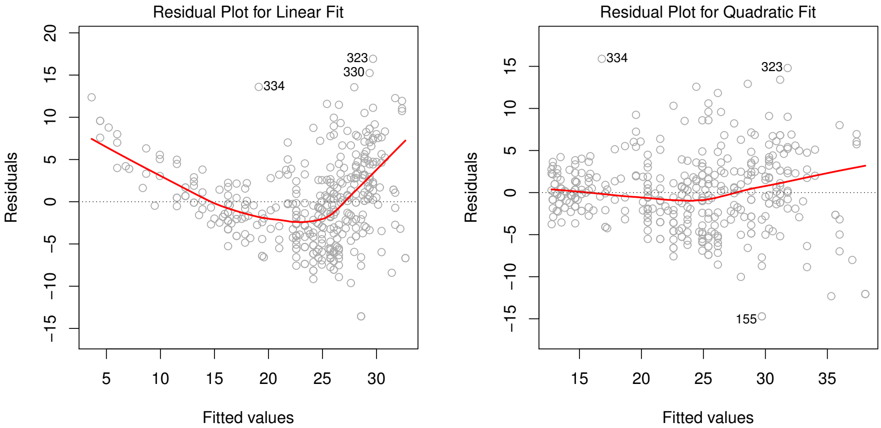
The quadratic term improves the fit.

Non-linear Relationships



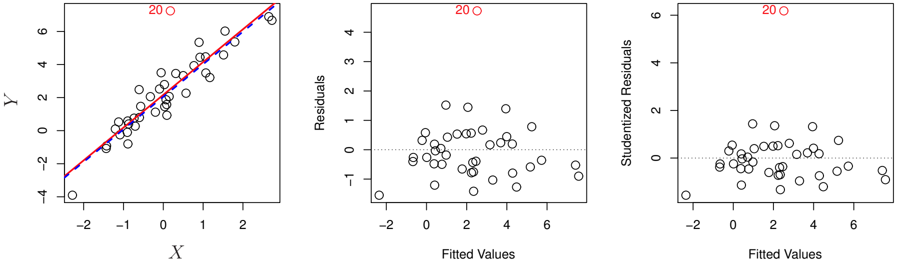
Remember the bias-variance trade-off.

Non-linear Relationships



A good way to spot non-linearities are *residual plots*.

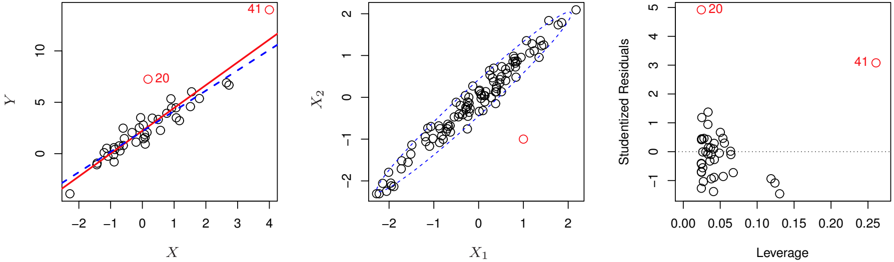
Outliers



- *Outliers* are data points with unusual response values y_i.
- They can be most easily spotted in residual plots.
- Even if they don't affect the parameters much they can badly influence statistics like R^2 and p-values.
- This can potentially change our interpretation of the model.

Be careful, only remove outliers for a good reason. In doubt report both fits.

High Leverage Points



- *High leverage points* are data points with unusual predictor values x_i.
- They tend to have a strong impact on the estimated parameters.
- For simple linear regression the *leverage* is computed like this:

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

In general, the h_i are the diagonal elements of the *hat matrix*.