Multiple Linear Regression

Rosalba Radice

Analytics Methods for Business

The Model

- Here our model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon.$$

- We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed. In the advertising example, the model becomes

$$\texttt{sales} = \beta_0 + \beta_1 \texttt{TV} + \beta_2 \texttt{radio} + \beta_4 \texttt{newspaper} + \epsilon.$$

Interpreting regression coefficients

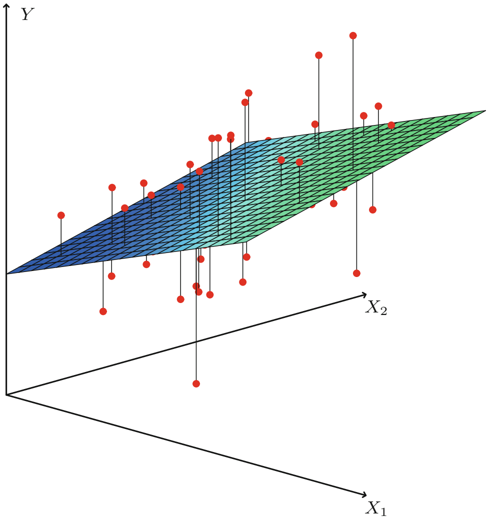
- The ideal scenario is when the predictors are uncorrelated:
  - Each coefficient can be estimated and tested separately.
  - Interpretations such as "a unit change in X_j is associated with a β_j change in Y, while all the other variables stay fixed", are possible.
- Correlations amongst predictors cause problems:
  - The variance of all coefficients tends to increase, sometimes dramatically.
  - Interpretations become hazardous - when X_j changes, everything else changes.
  - Claims of causality should be avoided for observational data.

Estimation and Prediction for Multiple Regression

- Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, we can make predictions using the formula
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p.$$
- We choose $\beta_0, \beta_1, \ldots, \beta_p$ to minimize the sum of squared residuals

$$\begin{aligned}
\text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\
&= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_p x_{ip} \right)^2
\end{aligned}$$

This is done using standard statistical software. The values $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ that minimize RSS are the multiple least squares regression coefficient estimates.



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.

Results for advertising data

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

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.

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Correlation matrix for `TV`, `radio`, `newspaper`, and `sales` for the `Advertising` data.

Some important questions

- Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
- Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Is at least one predictor useful?

- We test the null hypothesis,

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

  versus the alternative

$$H_a\text{: at least one } \beta_j \text{ is non-zero.}$$

- This hypothesis test is performed by computing the F-statistic,

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

Quantity	Value
Residual standard error	1.69
R^2	0.897
F-statistic	570

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.

Deciding on the important variables

- The most direct approach is called all subsets or best subsets regression: we compute the least squares fit for all possible subsets and then choose between them based on some criterion.
- However we often can't examine all possible models, since they are 2^p of them; for example when $p = 40$ there are over a billion models! Instead we need an automated approach that searches through a subset of them. We discuss two commonly used approaches.

Forward selection

- Begin with the null model - a model that contains an intercept but no predictors.
- Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.
- Add to that model the variable that results in the lowest RSS amongst all two-variable models.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

Backward selection

- Start with all variables in the model.
- Remove the variable with the largest p-value - that is, the variable that is the least statistically significant.
- The new $(p-1)$-variable model is fit, and the variable with the largest p-value is removed.
- Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value defined by some significance threshold.

Model Fit

- Two of the most common numerical measures of model fit are the RSE and R^2, where

$$\text{RSE} = \sqrt{\frac{1}{(n-p-1)}\text{RSS}},$$

- It turns out that R^2 will always increase when more variables are added to the model, even if those variables are only weakly associated with the response.
- Adjusted R^2 is a better option:

$$\bar{R}^2 = 1 - \frac{\text{RSS}/(n-p-1)}{\text{TSS}/(n-1)}$$

Prediction

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p$$

There are at least two sorts of uncertainty associated with this prediction. $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are estimates for $\beta_0, \beta_1, \ldots, \beta_p$. That is, the least squares plane

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_p X_p$$

is only an estimate for the true population regression plane

$$\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p.$$

The inaccuracy in the coefficient estimates is related to the *reducible error*. We can compute a confidence interval in order to determine how close $\hat{Y}$ will be to $\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$.

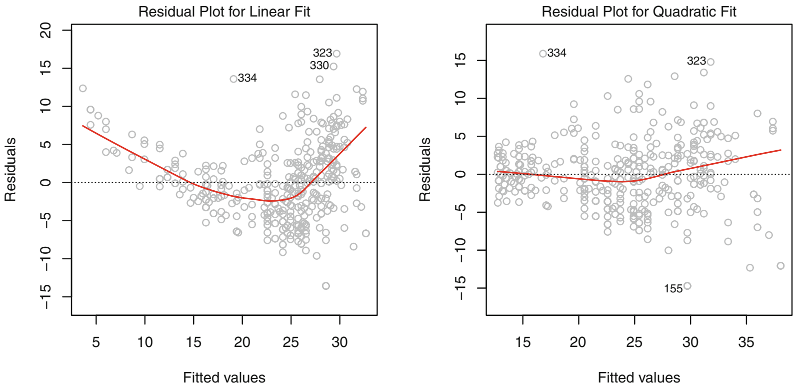
Prediction - continued

- Even if we knew the model - that is, even if we knew the true values for $\beta_0, \beta_1, \ldots, \beta_p$ - the response value cannot be predicted perfectly because of the random error ϵ. We refer to this as the *irreducible error*. How much will Y vary from $\hat{Y}$? We use prediction intervals to answer this question.
- Prediction intervals are always wider than confidence intervals.
- For example, given that \$100,000 is spent on `TV` advertising and \$20,000 is spent on `radio` advertising in each city, the 95% confidence interval is [10,985, 11,528].
  Given that \$100,000 is spent on `TV` advertising and \$20,000 is spent on `radio` advertising in that city the 95% prediction interval is [7,930, 14,580].

Potential Problems

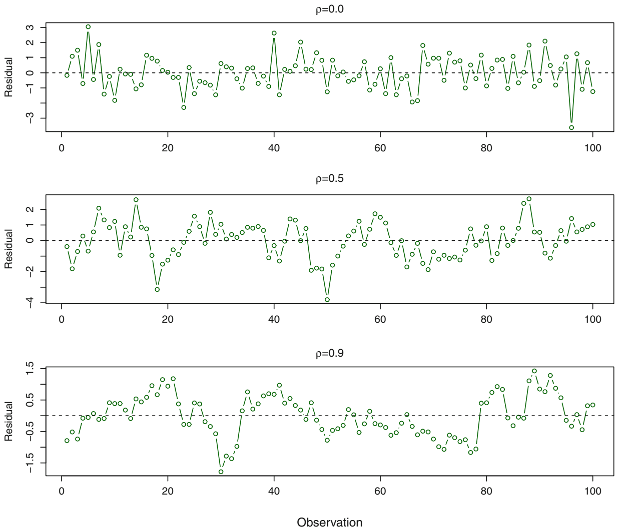
- Non-linearity of the response-predictor relationships
- Correlation of error terms
- Non-constant variance of error terms
- Outliers
- High-leverage points
- Collinearity

Non-linearity of the Data



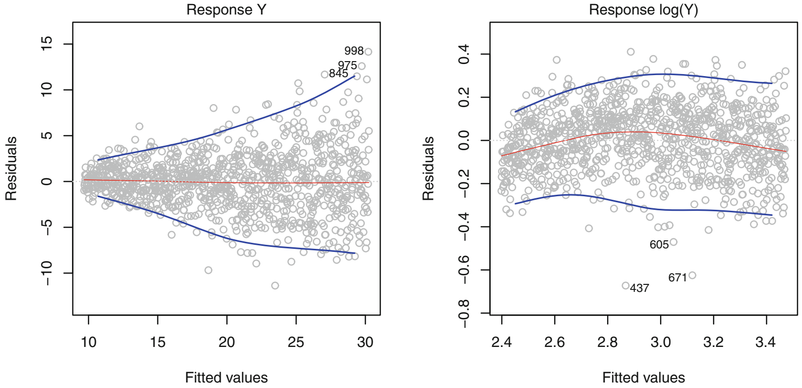
Plots of residuals versus predicted (or fitted) values for the `Auto` data set. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. Left: A linear regression of `mpg` on `horsepower`. A strong pattern in the residuals indicates non-linearity in the data. Right: A linear regression of `mpg` on `horsepower` and `horsepower`^{2}. There is little pattern in the residuals.

Correlation of Error Terms



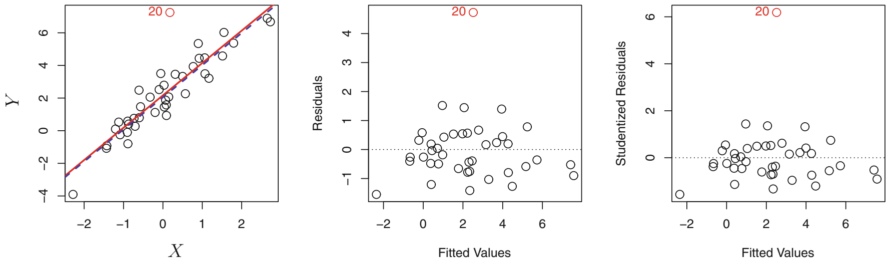
Plots of residuals from simulated time series data sets generated with differing levels of correlation ρ between error terms for adjacent time points.

Non-constant Variance of Error Terms



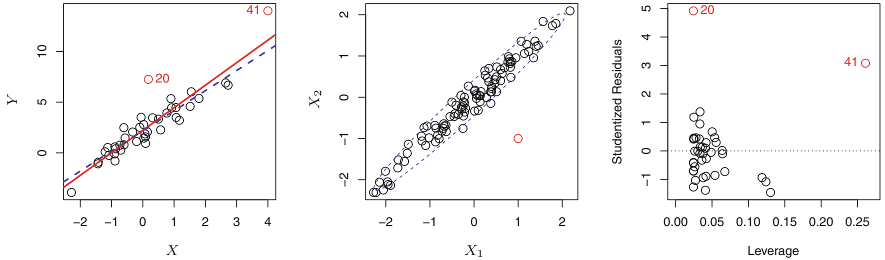
Residual plots. In each plot, the red line is a smooth fit to the residuals, intended to make it easier to identify a trend. The blue lines track the outer quantiles of the residuals, and emphasize patterns. Left: The funnel shape indicates heteroscedasticity. Right: The response has been log transformed, and there is now no evidence of heteroscedasticity.

Outliers



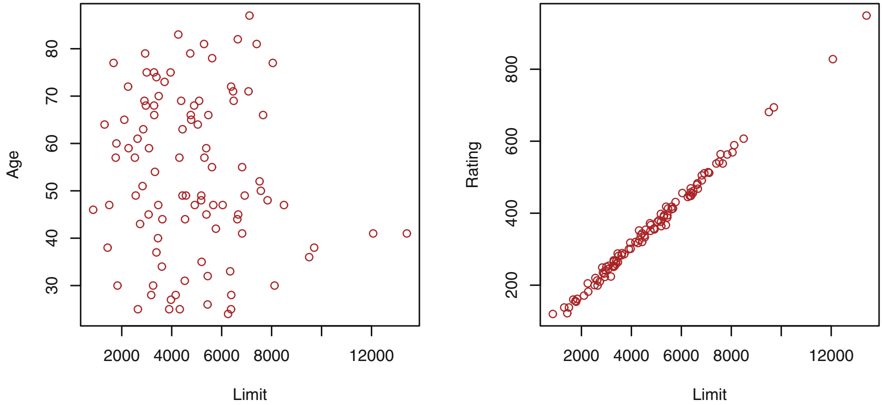
Left: The least squares regression line is shown in red, and the regression line after removing the outlier is shown in black. Center: The residual plot clearly identifies the outlier. Right: The outlier has a studentized (standardized) residual of 6; typically we expect values between 3 and 3.

High Leverage Points



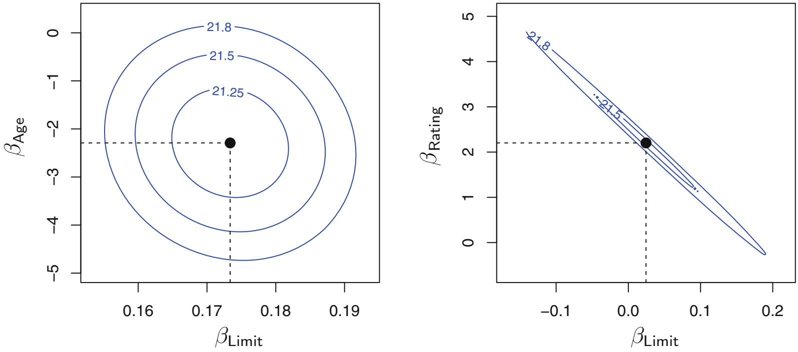
Left: Observation 41 is a high leverage point, while 20 is not. The red line is the fit to all the data, and the black line is the fit with observation 41 removed. Center: The red observation is not unusual in terms of its X_1 value or its X_2 value, but still falls outside the bulk of the data, and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

Collinearity



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

Collinearity - continued



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 $(\beta_{\text{Limit}}, \beta_{\text{Rating}})$ with a similar value for RSS.

Collinearity - continued

- A simple way to detect collinearity is to look at the correlation matrix of the predictors.
- Not all collinearity problems can be detected by inspection of the correlation matrix.
- A better way to assess multicollinearity is to compute the variance inflation factor (VIF)
- $$\text{VIF}\left(\hat{\beta}_j\right) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

  where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.
- If $R^2_{X_j|X_{-j}}$ is close to one, then collinearity is present, and so the VIF will be large.
- As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity.