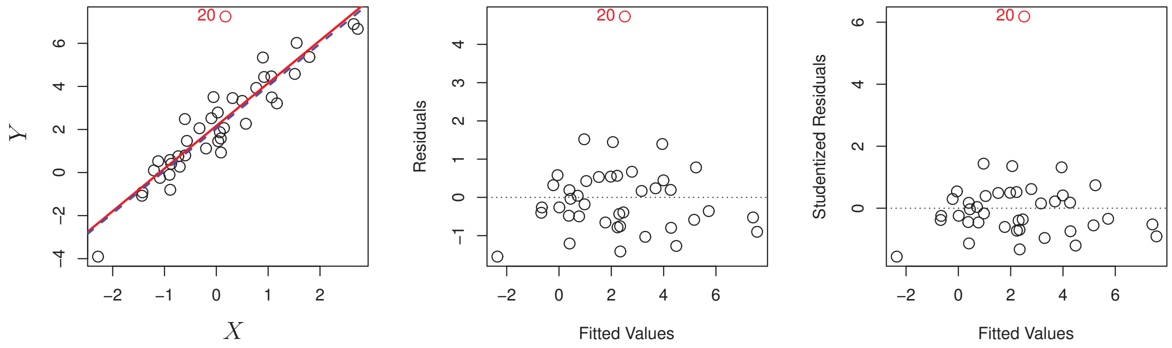
Outliers

- An outlier is a point for which y_i is far from the value predicted by the model. Outliers can arise for a variety of reasons, such as incorrect recording of an observation during data collection.
- It is typical for an outlier that does not have an unusual predictor value to have little effect on the least squares fit.
- However, even if an outlier does not have much effect on the least squares fit, it can cause other problems such as a dramatic increase in RSE or a dramatic decrease in R^2.



Left: The least squares regression line is shown in red, and the regression line after removing the outlier 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.

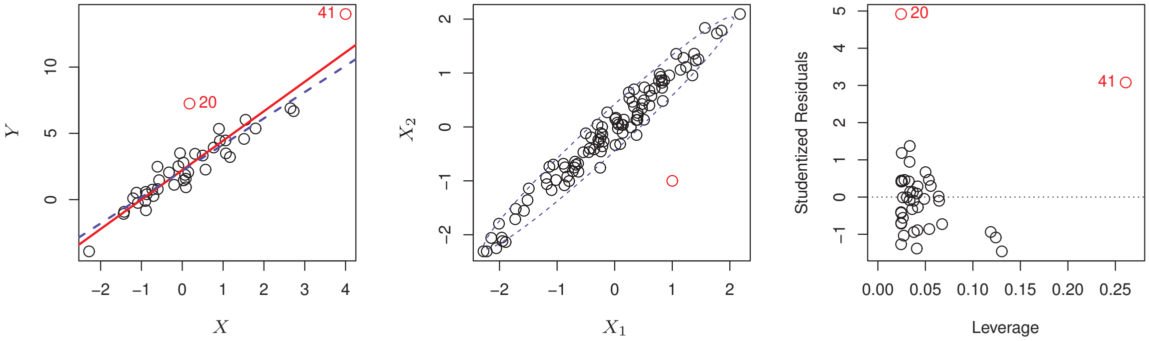
- Residual plots can be used to identify outliers. But in practice, it can be difficult to decide how large a residual needs to be before we consider the point to be an outlier.
- To address this problem, instead of plotting the residuals, we can plot the studentized residuals, computed by dividing each residual e_i by its estimated standard error.

In R, the function rstudent() will return the studentized residuals. Look at page 114 in the textbook.

- Observations whose studentized residuals are greater than 3 in absolute value are possible outliers.
- If we believe that an outlier has occurred due to an error in data collection or recording, then one solution is to simply remove the observation.

High Leverage Points

- We just saw that outliers are observations for which the response y_i is unusual given the predictor x_i. In contrast, observations with high leverage have an unusual value for x_i.
- It is cause for concern if the least squares line is heavily affected by just a couple of observations, because any problems with these points may invalidate the entire fit. For this reason, it is important to identify high leverage observations.



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.

- In a simple linear regression, high leverage observations are fairly easy to identify, since we can simply look for observations for which the predictor value is outside of the normal range of the observations.
- But in a multiple linear regression with many predictors, it is possible to have an observation that is well within the range of each individual predictor's values, but that is unusual in terms of the full set of predictors.
- In order to quantify an observation's leverage, we compute the *leverage statistic*. A large value of this statistic indicates an observation with high leverage.
- For a simple linear regression,

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

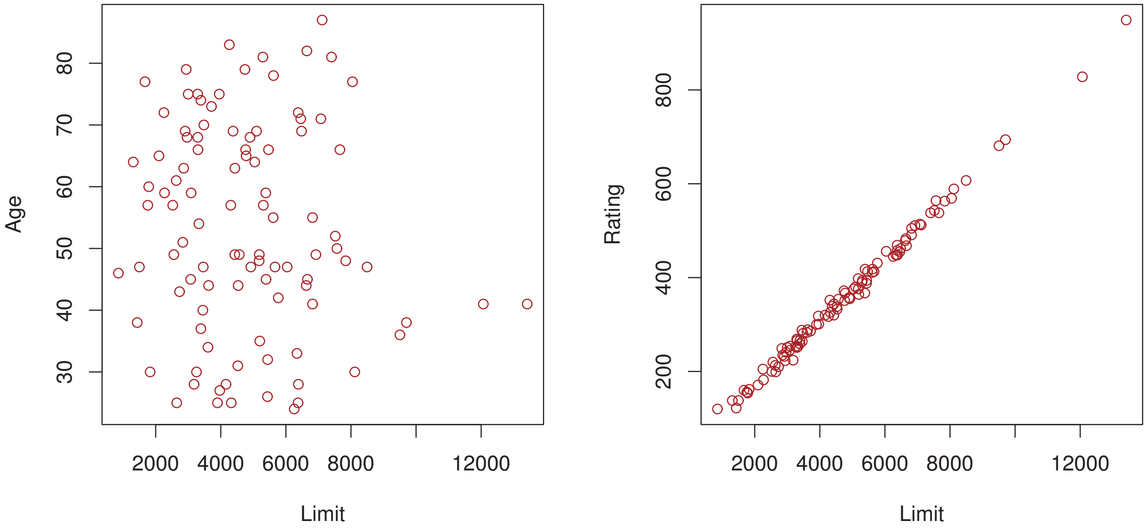
It is clear from this equation that h_i increases with the distance of x_i from $\bar{x}$.

- The leverage statistic h_i is always between $1/n$ and 1, and the average leverage for all the observations is always equal to $(p+1)/n$.
- So if a given observation has a leverage statistic that greatly exceeds $(p+1)/n$, then we may suspect that the corresponding point has high leverage.

In R, leverage statistics can be computed for any number of predictors using the hatvalues() function. Look at page 114 in the textbook.

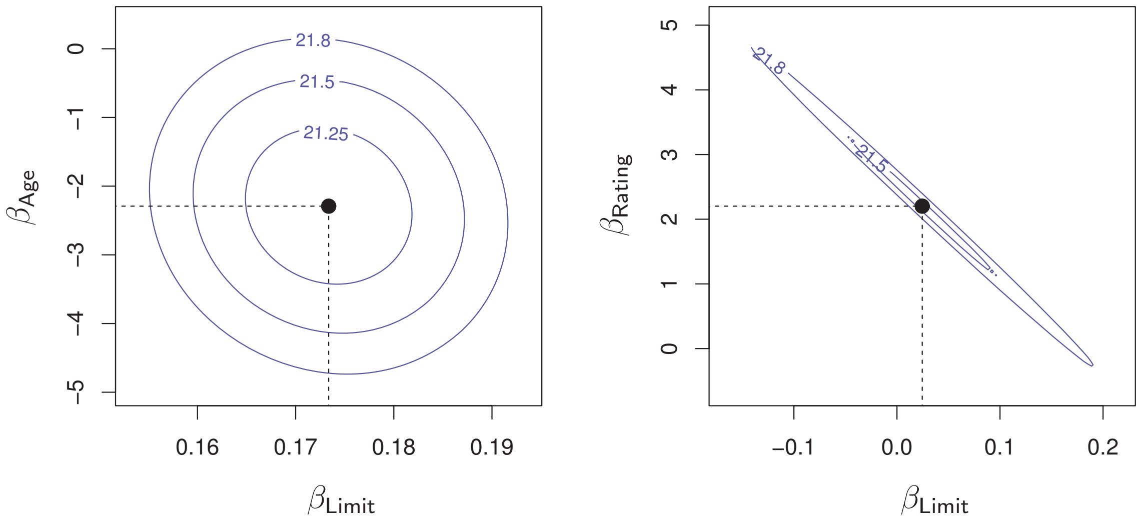
Collinearity

- Collinearity refers to the situation in which two or more predictor variables are closely related to one another.



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.

- The presence of collinearity can pose problems in the regression context, since it can be difficult to separate out the individual effects of collinear variables on the response.



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.

In the right panel, the contours run along a narrow valley. A small change in the data could cause the pair of coefficient values that yield the smallest RSS, that is, the least squares estimates, to move anywhere along this valley. This results in a great deal of uncertainty in the coefficient estimates.

- Since collinearity reduces the accuracy of the estimates of the regression coefficients, it causes the standard error for $\hat{\beta}_j$ to grow. Consequently, collinearity results in a decline in the t-statistic.
- As a result, in the presence of collinearity, we may fail to reject $H_0 : \beta_j = 0$. This means that the power of the hypothesis test—the probability of correctly detecting a non-zero coefficient—is reduced by collinearity.

		Coefficient	Std. error	t-statistic	p-value
Model 1	`Intercept`	−173.411	43.828	−3.957	< 0.0001
	`age`	−2.292	0.672	−3.407	0.0007
	`limit`	0.173	0.005	34.496	< 0.0001
Model 2	`Intercept`	−377.537	45.254	−8.343	< 0.0001
	`rating`	2.202	0.952	2.312	0.0213
	`limit`	0.025	0.064	0.384	0.7012

The results for two multiple regression models involving the Credit data set are shown. Model 1 is a regression of balance on age and limit, and Model 2 a regression of balance on rating and limit. The standard error of $\hat{\beta}_{\text{limit}}$ increases 12-fold in the second regression, due to collinearity.

In other words, the importance of the limit variable has been masked due to the presence of collinearity.

- A simple way to detect collinearity is to look at the correlation matrix of the predictors. An element of this matrix that is large in absolute value indicates a pair of highly correlated variables, and therefore a collinearity problem in the data.
- It is possible for collinearity to exist between three or more variables even if no pair of variables has a particularly high correlation. We call this situation *multicollinearity*.

- Instead of inspecting the correlation matrix, a better way to assess multicollinearity is to compute the variance inflation factor (VIF). The smallest possible value for VIF is 1, which indicates the complete absence of collinearity. A VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity.

- The VIF for each variable can be computed using the formula

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

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.

In R, the vif() function, part of the car package, can be used to compute variance inflation factors. Look at page 115 in the textbook.

- In the Credit data, a regression of balance on age, rating, and limit indicates that the predictors have VIF values of 1.01, 160.67, and 160.59.

- When faced with the problem of collinearity, there are two simple solutions.

  - The first is to drop one of the problematic variables from the regression.
  - The second solution is to combine the collinear variables together into a single predictor.

Comparison of Linear Regression with K-Nearest Neighbors

- As discussed before, linear regression is an example of a parametric approach because it assumes a linear functional form for $f(X)$.
- Parametric methods have several advantages.
  - They are often easy to fit, because one needs to estimate only a small number of coefficients.
  - In the case of linear regression, the coefficients have simple interpretations, and tests of statistical significance can be easily performed.
- But parametric methods do have a disadvantage: by construction, they make strong assumptions about the form of $f(X)$. If the specified functional form is far from the truth, and prediction accuracy is our goal, then the parametric method will perform poorly.
- In contrast, non-parametric methods do not explicitly assume a parametric form for $f(X)$, and thereby provide an alternative and more flexible approach for performing regression.
- Here we consider one of the simplest and best-known non-parametric methods, K-nearest neighbors regression (KNN regression).
- Given a value for K and a prediction point x_0, KNN regression first identifies the K training observations that are closest to x_0, represented by $\mathcal{N}_0$.
- It then estimates $f(x_0)$ using the average of all the training responses in $\mathcal{N}_0$. In other words,

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in N_0} y_i$$