# Regression Methods

**Stephen Jun Villejo**

School of Statistics

# Linear Regression

- Linear Regression is a simple supervised learning tool, and is useful for predicting a quantitative response.
- A *simple linear regression* assumes an approximately linear model between a quantitative response Y on the basis of a single predictor variable X.
- The model is given by:

$$Y \approx \beta_0 + \beta_1 X.$$

- More precisely, we have the model

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

- The systematic part of the model is $\beta_0 + \beta_1 X.$
- The term $\epsilon$ is a mean-zero random error term.

# Linear Regression

## Example

- For example, to answer the question if TV advertising is linearly related with sales, we will fit the linear model given by:

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}.$$

- $\beta_0$ and $\beta_1$ are two unknown constants that represent the intercept and slope terms, respectively. Both are called the model *coefficients* or *parameters.*
- Once we have used our data to produce estimates $\hat{\beta}_0$and $\hat{\beta}_1$, we can predict sales on the basis of a particular value of TV advertising through

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

# Linear Regression

## Estimating the Coefficients

- Our goal is to find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the resulting line is as close as possible to the data points.
- There are a number of ways of measuring *closeness*.
- The most common approach involves minimizing the *least squares* criterion.
- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the predicted value of Y for *i*th observation.
- Define $e_i = y_i - \hat{y}_i$ as the residual of the *i*th observation.

# Linear Regression

## Estimating the Coefficients

- The **least squares approach** chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ that will minimize

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

- The least squares estimators are:

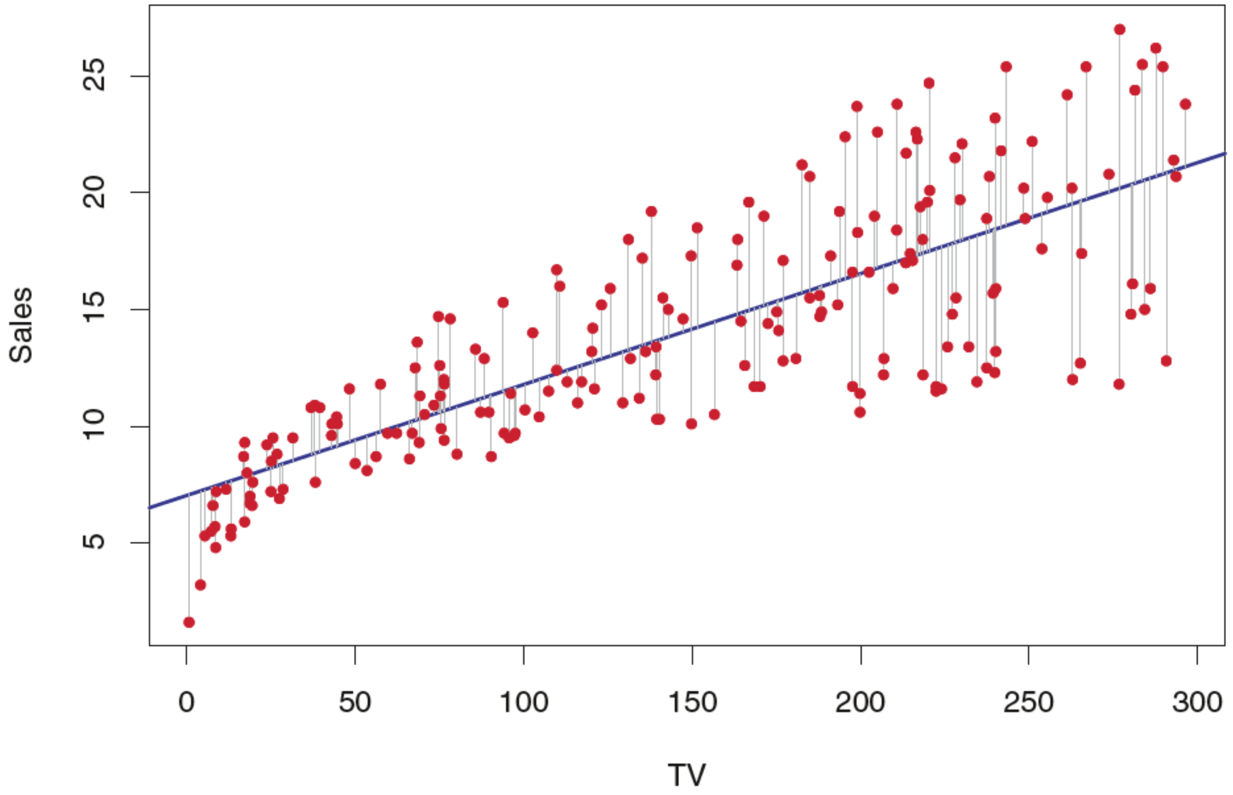
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

# Linear Regression

## Estimating the Coefficients

- The figure below displays the simple linear regression fit where $\hat{\beta}_0 = 7.03$ and $\hat{\beta}_1 = 0.0475$.



# Linear Regression

## Standard Errors

- How accurate are $\hat{\beta}_0$ and $\hat{\beta}_1$ as estimates for $\beta_0$ and $\beta_1$?
- We answer this by computing the standard error of our estimates.
- Roughly speaking, the standard error tells us the average amount that the estimate differs from the actual value.

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

# Linear Regression

## Confidence Intervals

- Standard error can be used to compute **confidence intervals**.
- A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter.
- There is approximately a 95% chance that the interval

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

  will contain the true value of $\beta_1$.
- Similarly, for $\beta_0$ we have $\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0)$.

# Linear Regression

## Tests of Significance

- Standard error can also be used to perform **hypothesis tests** on the coefficients.
- The most common hypothesis is:

  $\mathrm{H_o}$: There is no linear relationship between X and Y.

  $\mathrm{H_1}$: There is a linear relationship between X and Y.
- Mathematically, this corresponds to testing:

  $\mathrm{H_o}: \beta_1 = 0$

  $\mathrm{H_1}: \beta_1 \neq 0$
- We look at the p-value of the test in making the decision whether to reject or not reject the null hypothesis.

# Linear Regression

## Tests of Significance

- Roughly speaking, we interpret the **p-value** as follows: a small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance.
- Thus, we *reject the null hypothesis* – that is, we declare a relationship to exist between X and Y – if the p-value is small enough.
- Typical cutoffs for rejecting the null hypothesis are 5% or 1%.

# Linear Regression

## Tests of Significance

- The computation of the p-value is based on the *t-statistic* given by

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

- In our Advertising example, we have:

| | Coefficient | Std. error | t-statistic | p-value |
|---|---|---|---|---|
| Intercept | 7.0325 | 0.4578 | 15.36 | $< 0.0001$ |
| TV | 0.0475 | 0.0027 | 17.67 | $< 0.0001$ |

# Linear Regression

## Assessing the Accuracy of the Model

- The quality of a linear regression fit is typically assessed using two related quantities: **residual standard error** and $\mathbf{R^2}$ statistic.
- The residual standard error (RSE) is an estimate of the standard deviation, $\epsilon$, of the linear regression model.
- It is computed using the formula:

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

- The RSE is considered a measure of the *lack of fit*.

# Linear Regression

## Assessing the Accuracy of the Model

- The $\mathbf{R^2}$ statistic takes the form of a proportion – the proportion of variance explained by the model – and so it always takes on a value between 0 and 1.
- To calculate $\mathbf{R^2}$, we have:

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

- TSS is the total sum of squares, defined as $\mathrm{TSS} = \sum(y_i - \bar{y})^2$
- The $\mathbf{R^2}$ statistic measures the amount of variability in Y that can be explained using X.

# Linear Regression

## Assessing the Accuracy of the Model

- The $\mathbf{R^2}$ statistic has an interpretational advantage over the RSE.
- However, it can still be challenging to determine what is a *good* $\mathrm{R}^2$ value.
- Recall the correlation between X and Y defined as

$$\mathrm{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

- The $\mathrm{R}^2$ is the square of $\mathrm{Cor}(\mathrm{X},\mathrm{Y})$.

# R Exercise

# Multiple Linear Regression

- In practice we have more than one predictors. Instead of fitting a separate simple linear regression model for each predictor, we extend the simple linear regression model to

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

- Here, we interpret $\beta_j$ as the average effect on Y of a one unit increase in $X_j$, holding all other predictors fixed.
- In our advertising example, we have

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

# Multiple Linear Regression

## Estimating the Regression Coefficients

- The parameters are also estimated using the same least squares approach.
- We choose $\beta_0, \beta_1, \ldots, \beta_p$ to minimize the sum of squared residuals

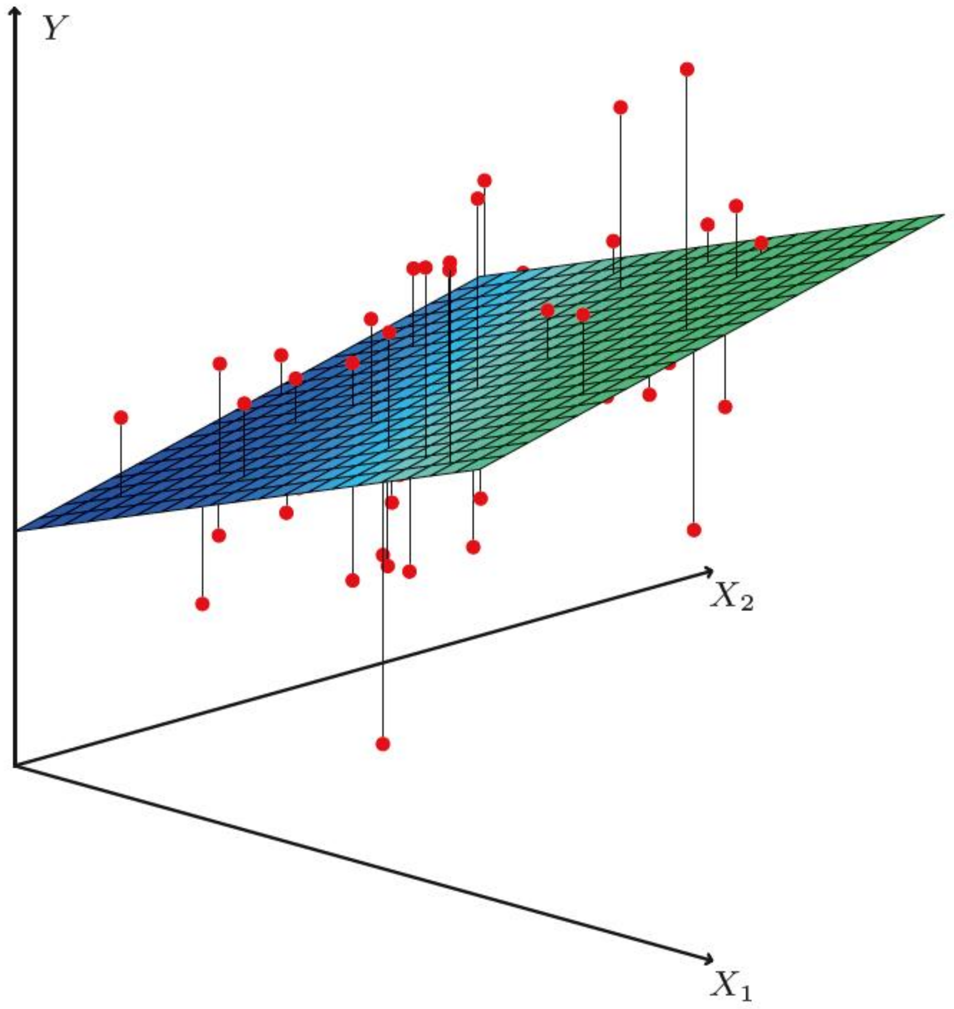
$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2. \end{aligned}$$

- Unlike the simple linear regression coefficient estimates, the multiple linear regression coefficient estimates have somewhat complicated forms that are most easily represented using matrix algebra.

# Multiple Linear Regression

## Estimating the Regression Coefficients

- 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 and the plane.



# Multiple Linear Regression

## Example

- For the Advertising data, we have the following coefficient estimates:

| | 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 |

# Multiple Linear Regression

## Some Important Questions

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

# Multiple Linear Regression

## Is There a Relationship Between the Response and Predictors?

- We use a hypothesis test to answer this question.
- We test the null hypothesis,

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

versus the alternative

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

- The hypothesis is performed by computing the *F-statistic*

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

# Multiple Linear Regression

## Is There a Relationship Between the Response and Predictors?

- When there is no relationship between the response and predictors, one would expect the F-statistics to take on a value close to 1.
- A large F-statistic suggests that at least one of the predictors is significantly related to the response.
- The p-value associated with the F-statistic is used to determine whether or not to reject $\mathrm{H_o}$.

# Multiple Linear Regression

## Is There a Relationship Between the Response and Predictors?

- We also test the **partial effect** of each variable in the mode.
- These provide information about whether each individual predictor is related to the response, after adjusting for the other predictors.
- Test on the partial effect of the individual coefficients uses the t-statistic.
- In our Advertising example, we have:

| | 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 |

# Multiple Linear Regression

## Remarks

- The approach of using an F-statistic to test for any association between the predictors and the response works when p is relatively small, and certainly small compared to n.
- If p > n, we cannot even fit the multiple linear regression model using least squars.
- When p is large, we use *high-dimensional* techniques.

# Multiple Linear Regression

## Deciding on Important Variables

- If we conclude that at least one of the predictors is related to the response, we want to know which are the *guilty* ones!
- Note that the RSS always decreases as more variables are added to the model.
- Determining which predictors (a subset) are associated with the response, in order to fit a single model, is referred to as ***variable selection***.
- For the many possible models, we determine the optimal model based on some criteria: Mallows's $C_p$, Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted $R^2$.

# Multiple Linear Regression

## Deciding on Important Variables

### Mallows's $C_p$

$$C_p = \frac{1}{n}\left(\text{RSS} + 2d\hat{\sigma}^2\right),$$

- The $C_p$ is an estimate of the MSE for a model with $d$ predictors. The $C_p$ adds a penalty to the RSS to adjust for the corresponding decrease in the RSS.
- We choose the model with the lowest $C_p$ value.

# Multiple Linear Regression

## Deciding on Important Variables

### AIC

$$\text{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\text{RSS} + 2d\hat{\sigma}^2\right)$$

- AIC is proportional to the $C_p$ value.

### BIC

$$\text{BIC} = \frac{1}{n}\left(\text{RSS} + \log(n)d\hat{\sigma}^2\right)$$

- Like the $C_p$ value and AIC, we select the model with the lowest BIC value.

# Multiple Linear Regression

## Deciding on Important Variables

### Adjusted $R^2$

$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}.$$

- We select the model with the largest adjusted $R^2$.
- The intuition behind the adjusted $R^2$ is that once all of the correct variables have been included in the model, adding additional *noise* variables will lead to only a very small decrease in RSS.
- The adjusted $R^2$ statistic pays a price for the inclusion of unnecessary variables.

# Multiple Linear Regression

## Deciding on Important Variables

- Ideally, we try out all possible subset of the predictors. However, trying out every possible subset may be infeasible, even for moderate p.
- We have automated and efficient approaches to choose a smallest set of models to consider using: **Forward selection**, **Backward selection**, **Mixed** or **Stepwise selection**

# Multiple Linear Regression

## Deciding on Important Variables

### Forward selection

- We begin with a null model.
- We then add to the model the variable that results in the lowest RSS.
- Then, we choose the next variable that will result in the lowest RSS for the new two-variable model.
- This is continued until some stopping rule is satisfied.

# Multiple Linear Regression

## Deciding on Important Variables

### Backward selection

- We start with all variables in the model then remove the variable with the largest p-value, that is, the variable that is the least significant.
- The new (p-1)-variable model is fit, and the variable with the largest p-value is removed.
- This continues until all remaining variables have a p-value below some threshold.

# Multiple Linear Regression

## Deciding on Important Variables

### Mixed selection

- This is a combination of forward and backward selection.

# Multiple Linear Regression

## Prediction

- The *least squares plane*

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

  is only an estimate for the true population regression plane

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

- We can compute a confidence interval to determine how close $\hat{Y}$ will be to $f(x)$.
- Even if we know the true values $\beta_0$, $\beta_1, \ldots, \beta_p$, the response value cannot be predicted perfectly because of the random error.
- We use **prediction intervals** to quantify how much $\hat{Y}$ will vary from Y.
- **Prediction intervals** are wider than confidence intervals.

# Other Considerations in the Regression Model

## Qualitative Predictors

- To include qualitative predictors, we make use of dummy or indicator variables.

### Predictors with Only Two Levels

- Here, we use only one indicator variable. For example, for *Gender:*

$$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male,} \end{cases}$$

- This results to the model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

- $\beta_1$ is the average difference in Y between females and males.

# Other Considerations in the Regression Model

## Qualitative Predictors

### Predictors with More Than Two Levels

- If a variable has d levels, we introduce (d-1) dummy variables.
- Example, for ethnicity = {Asian, Caucasian, African American} :

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

- Then we have the model:

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

# Other Considerations in the Regression Model

## Qualitative Predictors

### Predictors with More Than Two Levels

- Then we have the model:

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

- $\beta_1$ can be interpreted as the difference in the average response between Asian and African American categories.
- $\beta_2$ is the average difference in the average response between Caucasian and African American.
- The level with no dummy variable is the **baseline** category.

# Other Considerations in the Regression Model

## Interaction Terms

- Previously, we concluded that both TV and radio seem to be associated with sales.
- We assumed that the effect of TV is independent of radio.
- Suppose that spending money on radio advertising actually increases the effectiveness of TV advertising, so that the slope term for TV should increase as radio increases?
- To account for this interaction, we include an **interaction term**.

# Other Considerations in the Regression Model

## Interaction Terms

- We have the model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

- We can rewrite this as

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

- The effect of $X_1$ is no longer constant: adjusting $X_2$ will change the impact of $X_1$ on Y.

# Other Considerations in the Regression Model

## Interaction Terms

- **Example**
- Suppose we wish to predict credit balance using income and whether student or not.
- Without interaction term, the model is

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

# Other Considerations in the Regression Model

## Interaction Terms

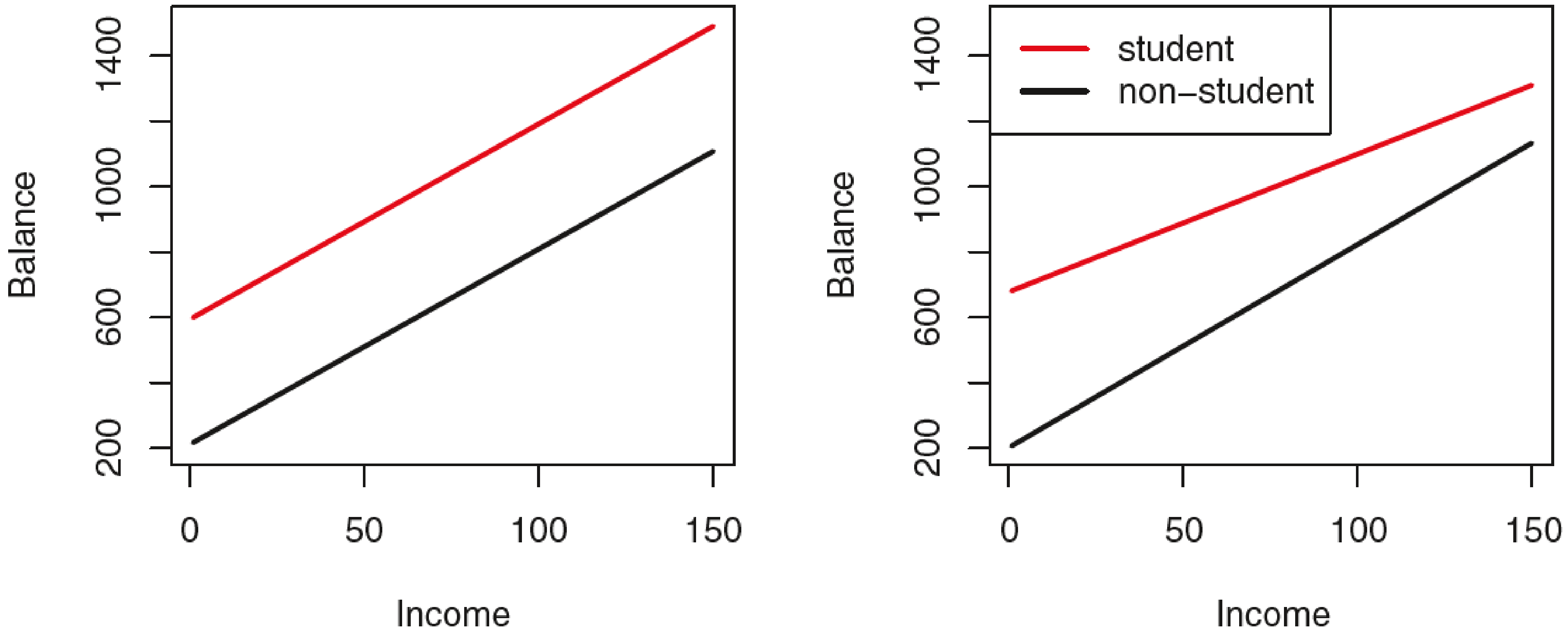
- **Example**
- Introducing the interaction term between student and income, we have:

$$
\begin{aligned}
\texttt{balance}_i &\approx \beta_0 + \beta_1 \times \texttt{income}_i + \begin{cases} \beta_2 + \beta_3 \times \texttt{income}_i & \text{if student} \\ 0 & \text{if not student} \end{cases} \\
&= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \texttt{income}_i & \text{if student} \\ \beta_0 + \beta_1 \times \texttt{income}_i & \text{if not student} \end{cases}
\end{aligned}
$$

# Other Considerations in the Regression Model

## Interaction Terms

- **Example**



## Potential Problems in the Model

**The most common problems in a linear regression model are:**

1. Non-linearity of the response-predictor relationships
2. Correlation of error terms
3. Non-constant variance of error terms
4. Outliers
5. High-leverage points
6. Collinearity

# Potential Problems in the Model

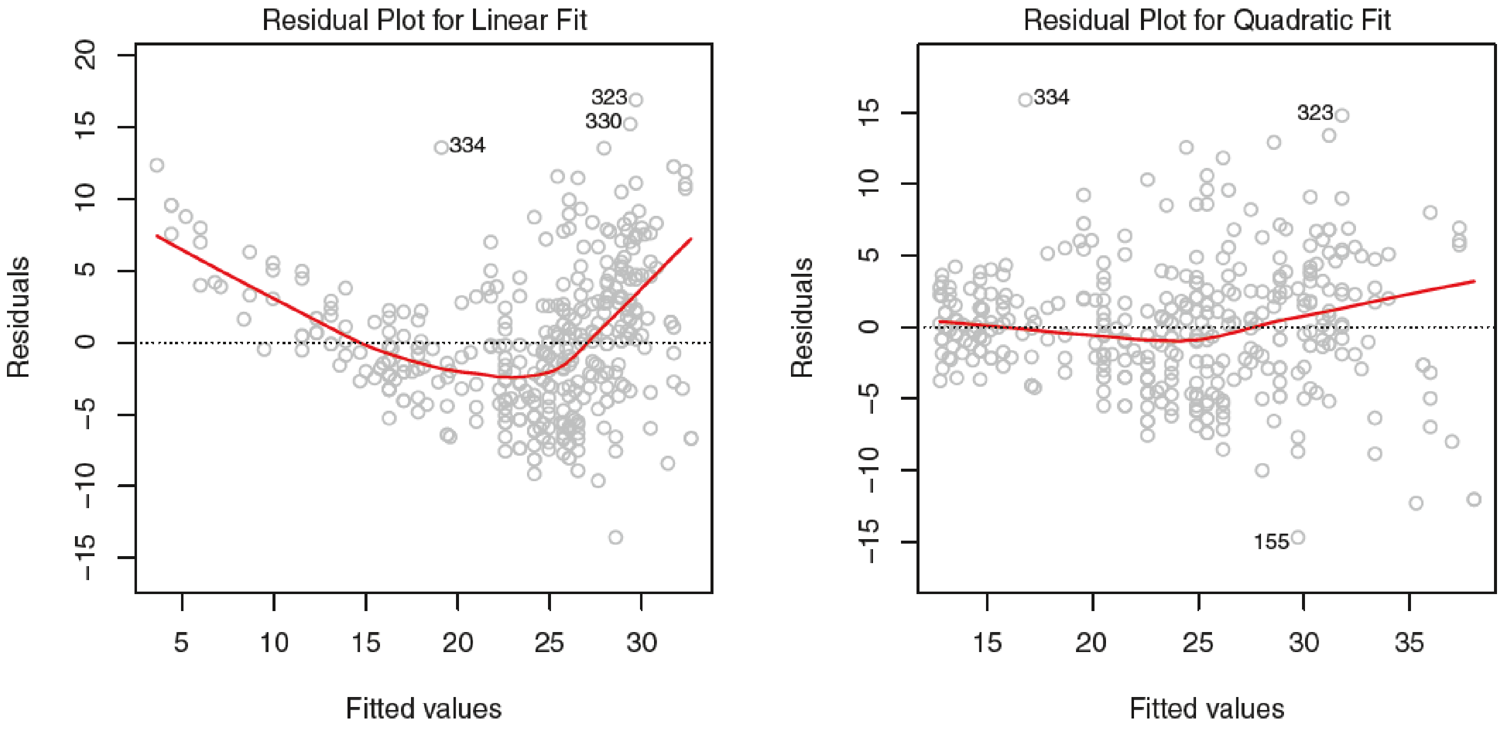
## Non-linearity

- If the true relationship is far from linear, then all conclusions we draw are suspect. In addition, the prediction accuracy can be significantly reduced.
- **Residual plots** are a useful graphical tool for identifying non-linearity.
- Ideally, the residual plot will not shown any discernible pattern if there is no non-linear relationship.
- If the residual plot indicates that there are non-linear associations in the data, then a simple approach is to use non-linear transformations of the predictors.

# Potential Problems in the Model

## Illustration

- Figure on the left shows a strong pattern of non-linearity.
- Figure on the right is a residual plot with quadratic terms.



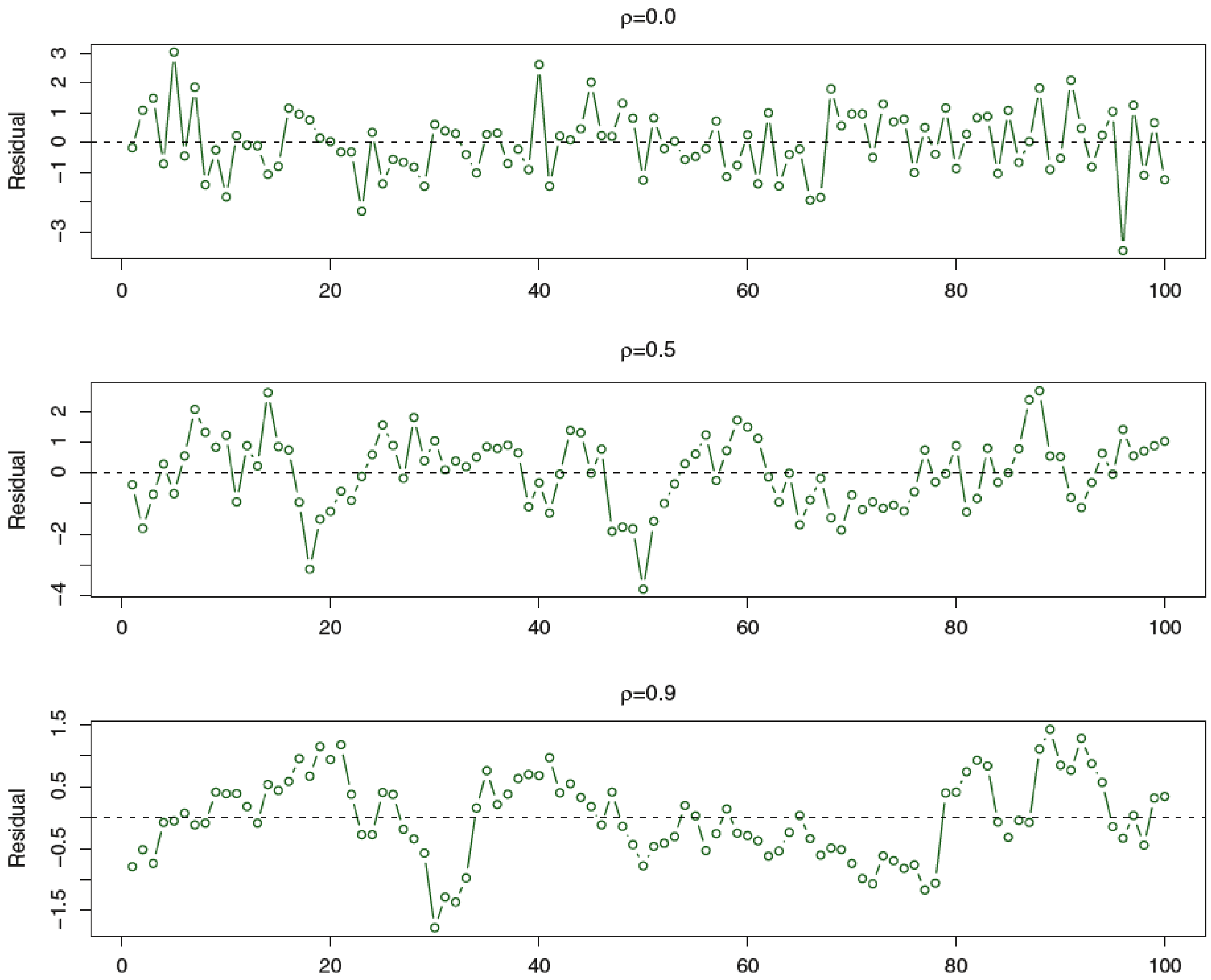
# Potential Problems in the Model

## Correlation of Error Terms

- An important assumption of the linear model is that the error terms are uncorrelated.
- If there is correlation in the error terms, then the estimated standard errors will tend to underestimate the true standard errors.
- Such correlations frequently occur in the context of *time series* data.

# Potential Problems in the Model

## Correlation of Error Terms



# Potential Problems in the Model

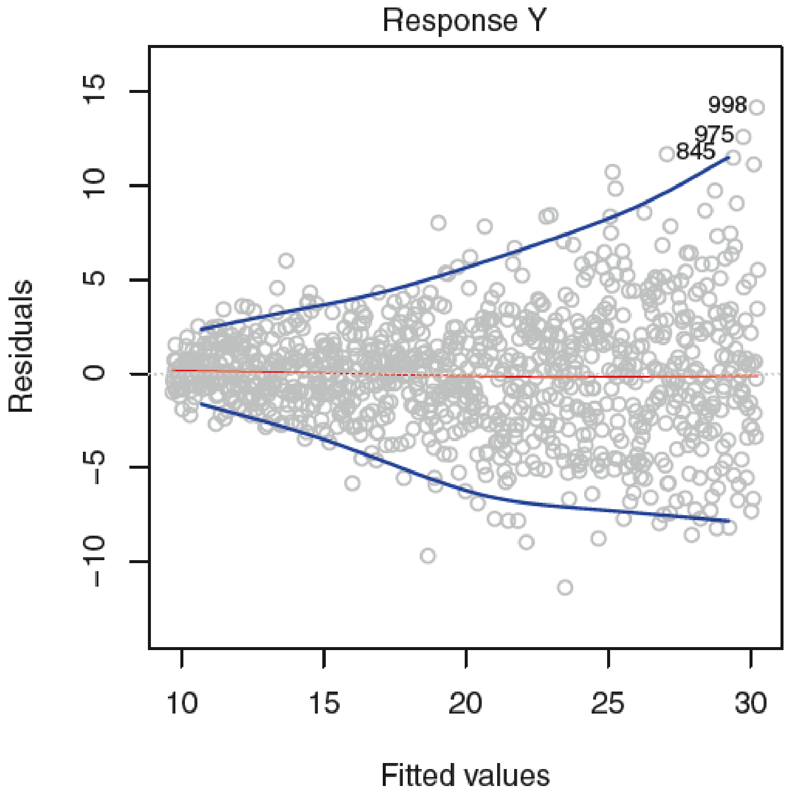
## Nonconstant Variance of Error Terms

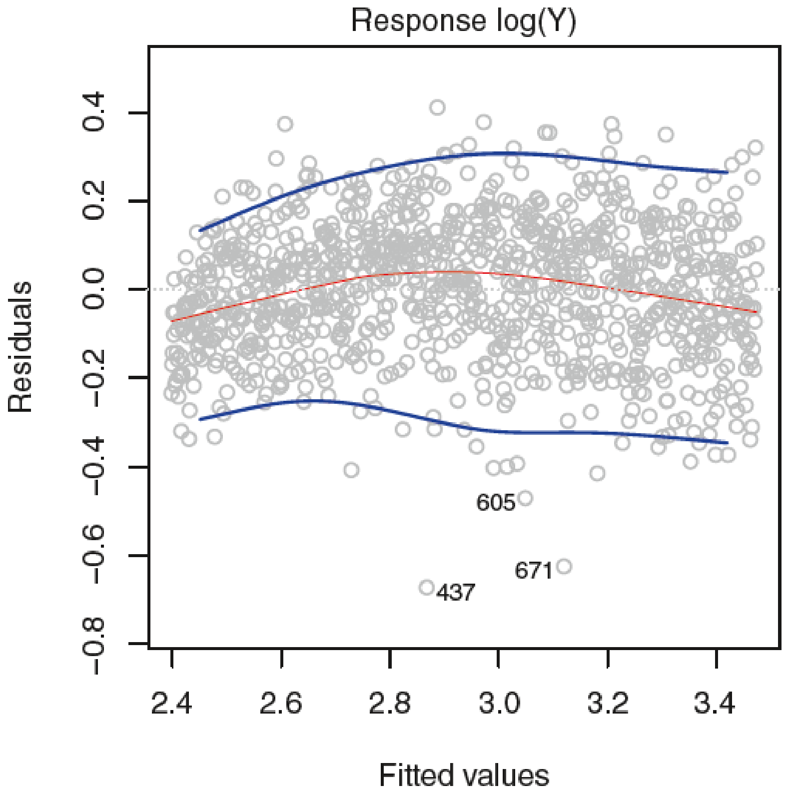
- Another assumption of the linear regression model is that the error term has constant variance.
- The problem of nonconstant variances is called *heteroscedasticity*.
- One can identify the presence of *heteroscedasticity* from the presence of a funnel shape in the residual plot.
- When faced with this problem, one possible solution is to transform the response Y using a concave function such as log(Y) or $\sqrt{Y}$.
- Another option is to use *weighted least squares*.

# Potential Problems in the Model

## Nonconstant Variance of Error Terms

- Residual plot on the left resembles a funnel shape indicating heteroscedasticity. The predictor has been log transformed on the right figure.

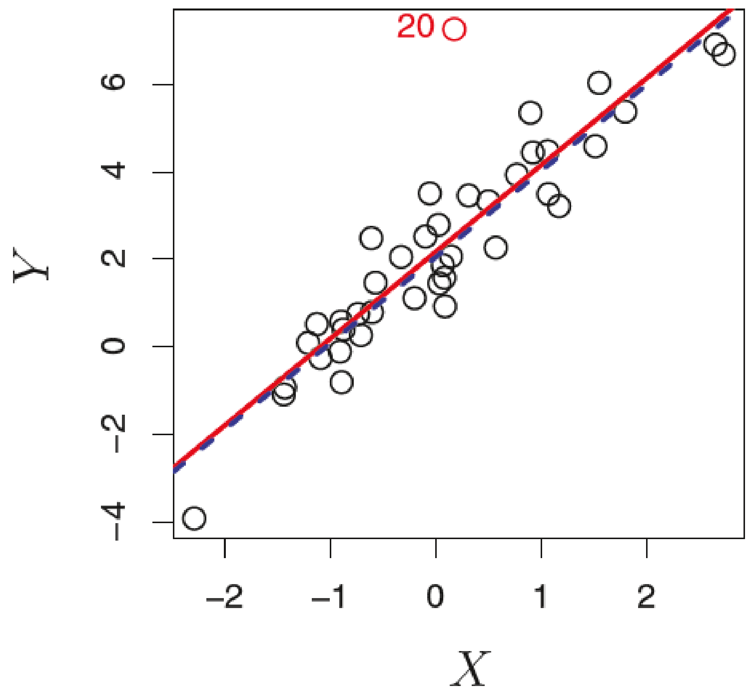


# Potential Problems in the Model

## 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.

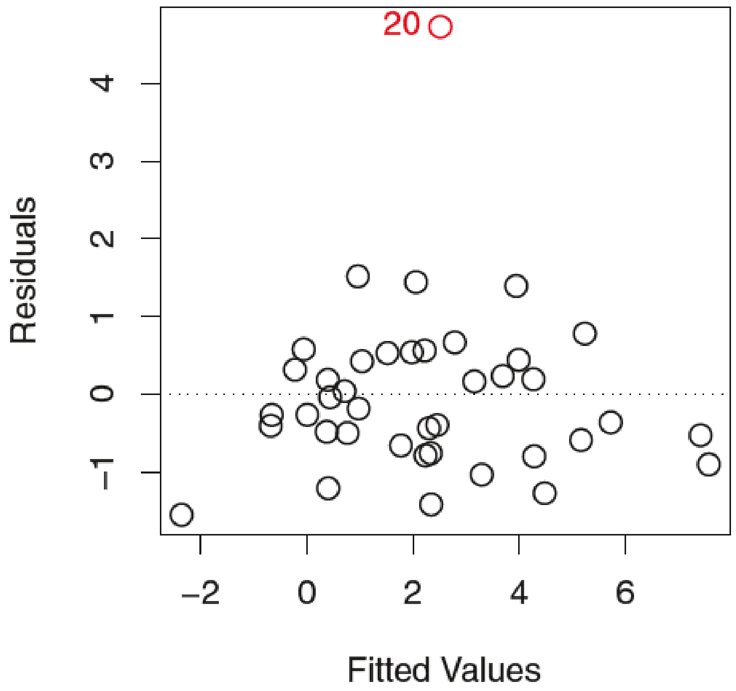


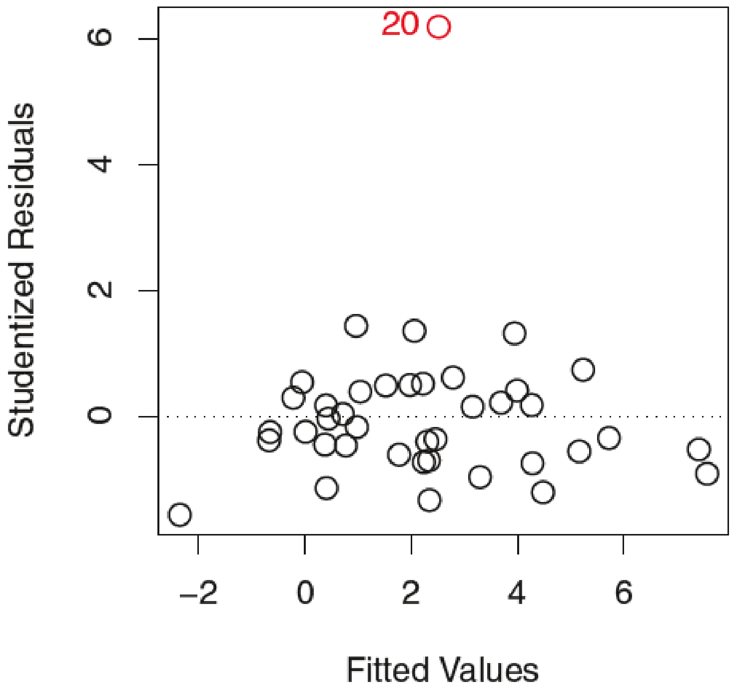

- The red point on the left shows a typical outlier.
- The blue dashed line is the least squares fit after removing the outlier.
- It is typical for an outlier that does not have an unusual predictor value to have little effect on the least squares fit.
- However, it can cause the $R^2$ fit to decline or the standard errors to increase.

# Potential Problems in the Model

## Outliers

- Residuals can be used to identify outliers.
- Observations whose studentized residuals are greater than 3 in absolute value are possible outliers.

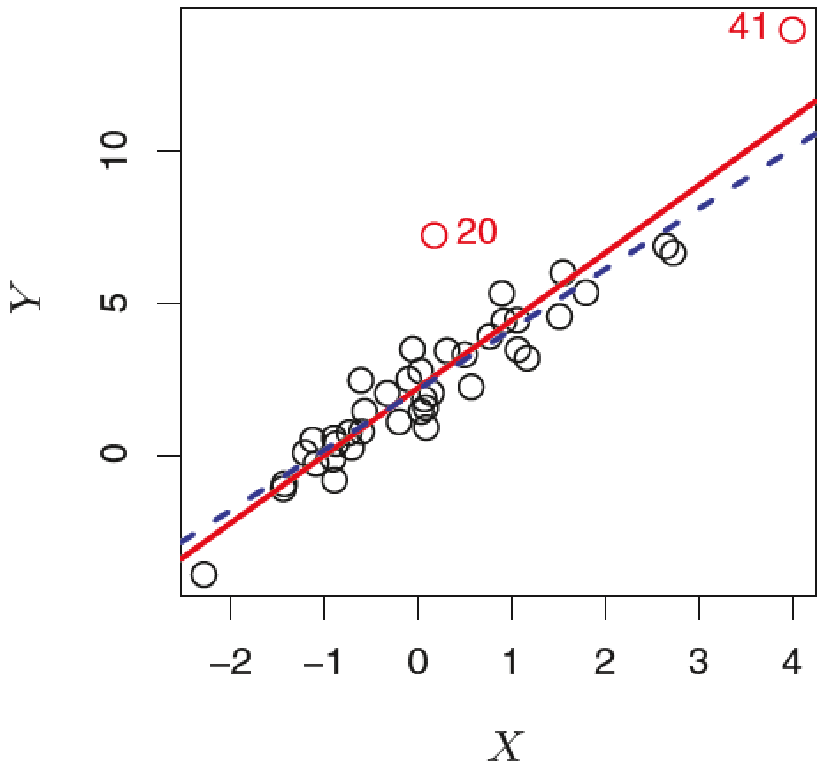


# Potential Problems in the Model

## High Leverage Points

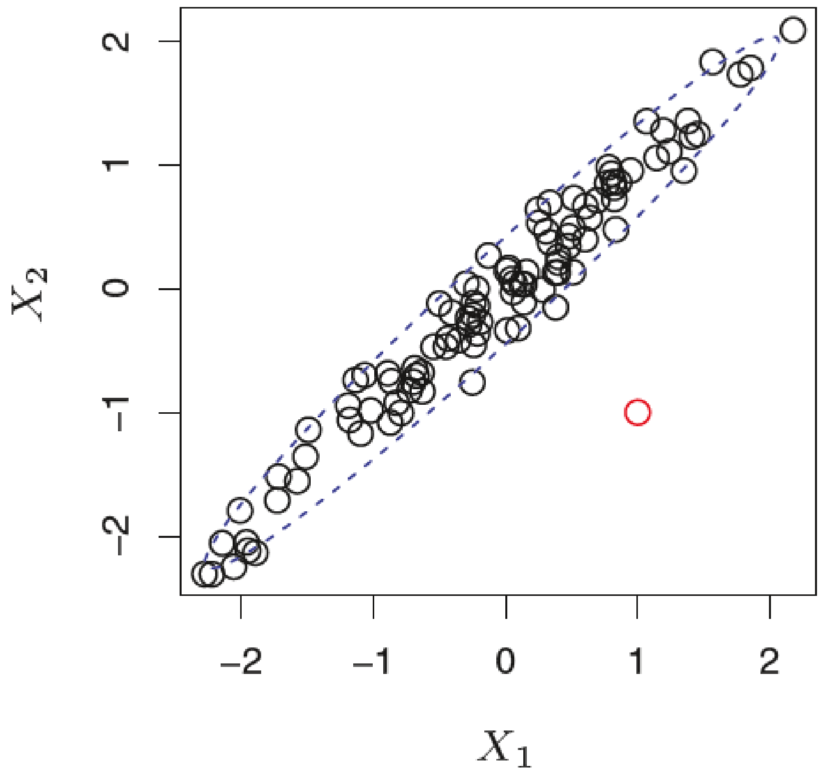
- Observations with *high leverage* have an unusual value for $x_i$.




- Observation 41 on the left has high leverage. Observation 20 has a small leverage.
- Removing an observation with high leverage has a more substantial impact on the least squares line.

# Potential Problems in the Model

## High Leverage Points



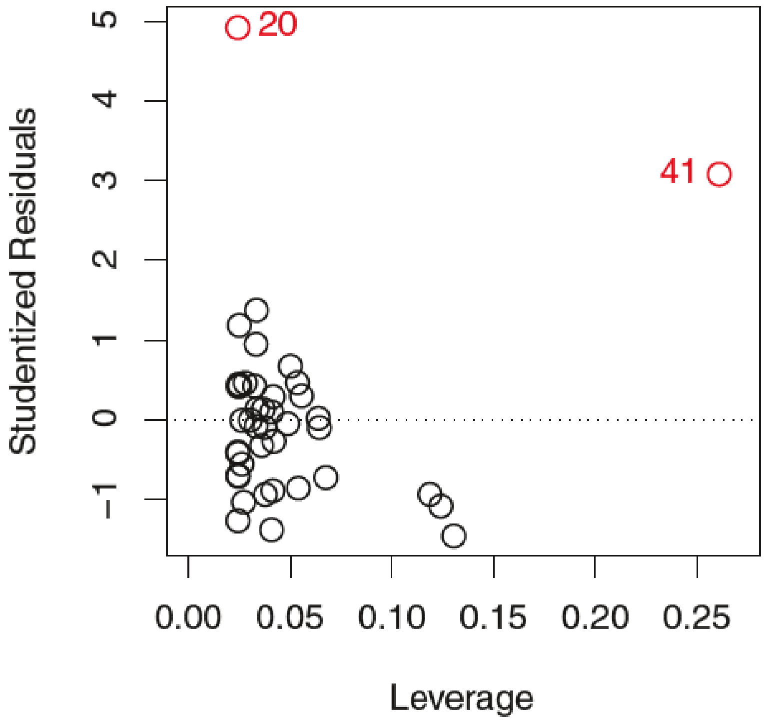

- The red observation is not unusual in terms of its $X_1$ value or its $X_2$ value, but still falls outside the bulk of the data, and hence has high leverage.
- This is a problem in multiple linear regression.

# Potential Problems in the Model

## High Leverage Points

- To quantify an observation's leverage, we compute the *leverage statistic* given by

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




- A given observation with leverage statistic that greatly exceeds $(p+1)/n$ has high leverage.

# Potential Problems in the Model

## Collinearity

- *Collinearity* refers to the situation in which two or more predictor variables are closely related to one another.
- 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.
- Collinearity reduces the accuracy of the estimates as it causes the standard error to grow.
- A simple way to detect collinearity is to look at the correlation matrix of the predictors.
- However, it is possible for collinearity to exist even if no pair of variables have high correlation.

# Potential Problems in the Model

## Collinearity

- The **Variance Inflation Factor (VIF)** is a better metric to assess multicollinearity. It is given by

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

- As a rule of thumb, a VIF that exceeds 5 or 10 indicates a problematic amount of collinearity.
- When faced with collinearity, common solutions are: drop one of the problematic variables, or combine them into a single predictor.

# R Exercise