Introduction to Machine Learning – Linear Regression Part 2

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Jordan Crouser (https://jcrouser.github.io/)

Plan for Today

- More considerations for regression modeling
  - Qualitative predictors
  - Extensions of linear model
  - Potential problems

Warm Up

- RSE (Residual Standard Error)
  - Estimate of the standard deviation of ϵ, or the average amount responses (in the data) will deviate from the regression line
  - $\sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{n-2}}$
  - Will a worse fit lead to a bigger or smaller RSE?
- R^2 statistic
  - Estimate of the proportion of variance explained by the model
  - $\frac{TSS - RSS}{RSS} = 1 - \frac{RSS}{TSS}$
  - What value of R^2 indicates the worst possible fit? What value indicates the best possible?

Qualitative Predictors

Motivation

So far, we have assumed all variables in our linear regression models are *quantitative*.

Let's look at a new dataset.

Motivation

`Carseats` dataset:

- **Description:** simulated data set on sales of car seats
- **Format:** 400 observations on the following 11 variables
  - **`Sales`**: unit sales at each location
  - **`CompPrice`**: price charged by nearest competitor at each location
  - **`Income`**: community income level
  - **`Advertising`**: local advertising budget for company at each location
  - **`Population`**: population size in region (in thousands)
  - **`Price`**: price charged for car seat at each site
  - **`ShelveLoc`**: quality of shelving location at site (Good | Bad | Medium)
  - **`Age`**: average age of the local population
  - **`Education`**: education level at each location
  - **`Urban`**: whether the store is in an urban or rural location
  - **`USA`**: whether the store is in the US or not

Motivation

`Carseats` dataset:

- **Description:** simulated data set on sales of car seats
- **Format:** 400 observations on the following 11 variables
  - **`Sales`**: unit sales at each location
  - **`CompPrice`**: price charged by nearest competitor at each location
  - **`Income`**: community income level
  - **`Advertising`**: local advertising budget for company at each location
  - **`Population`**: population size in region (in thousands)
  - **`Price`**: price charged for car seat at each site
  - **`ShelveLoc`**: quality of shelving location at site (Good | Bad | Medium)
  - **`Age`**: average age of the local population
  - **`Education`**: education level at each location
  - **`Urban`**: whether the store is in an urban or rural location
  - **`USA`**: whether the store is in the US or not

What if I want to use Urban to predict Sales?

Motivation

`Carseats` dataset:

- **Description:** simulated data set on sales of car seats
- **Format:** 400 observations on the following 11 variables
  - **`Sales`**: unit sales at each location
  - **`CompPrice`**: price charged by nearest competitor at each location
  - **`Income`**: community income level
  - **`Advertising`**: local advertising budget for company at each location
  - **`Population`**: population size in region (in thousands)
  - **`Price`**: price charged for car seat at each site
  - **`ShelveLoc`**: quality of shelving location at site (Good | Bad | Medium)
  - **`Age`**: average age of the local population
  - **`Education`**: education level at each location
  - **`Urban`**: whether the store is in an urban or rural location
  - **`USA`**: whether the store is in the US or not

What if I want to use Urban to predict Sales?

Can we do this with our current methods? Why or why not?

Motivation

So far, we have assumed all variables in our linear regression models are *quantitative*.

Sometimes, our predictors are *qualitative*.

Ex.

`Urban`: whether the store is in an urban or rural location

Two Levels

So far, we have assumed all variables in our linear regression models are *quantitative*.

Sometimes, our predictors are *qualitative*.

Ex.

`Urban`: whether the store is in an urban or rural location

Notice `Urban` only has 2 levels – urban or rural

To incorporate it into a regression, we create a *dummy variable*.

Two Levels

*A **dummy variable*** is an indicator variable that takes on two possible numerical values.

Ex. For Urban, our dummy variable will be: *X* where

$$x_i = \begin{cases} 1 & if\ ith\ sale\ was\ urban \\ 0 & if\ the\ ith\ sale\ was\ rural \end{cases}$$

Two Levels

A ***dummy variable*** is an indicator variable that takes on two possible numerical values.

Ex. For Urban, our dummy variable will be: *X* where

$$x_i = \begin{cases} 1 & if\ ith\ sale\ was\ urban \\ 0 & if\ the\ ith\ sale\ was\ rural \end{cases}$$

This is also called "one-hot encoding"

Two Levels

A ***dummy variable*** is an indicator variable that takes on two possible numerical values.

Ex. For Urban, our dummy variable will be: *X* where

$$x_i = \begin{cases} 1 & if\ ith\ sale\ was\ urban \\ 0 & if\ the\ ith\ sale\ was\ rural \end{cases}$$

Now, we can use X as a predictor in our regression equation

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & if\ the\ ith\ sale\ was\ urban \\ \beta_0 + \epsilon_i & if\ the\ ith\ sale\ was\ rural \end{cases}$$

Two Levels

*A **dummy variable*** is an indicator variable that takes on two possible numerical values.

Ex. For Urban, our dummy variable will be: *X* where

$$x_i = \begin{cases} 1 & if\ ith\ sale\ was\ urban \\ 0 & if\ the\ ith\ sale\ was\ rural \end{cases}$$

Now, we can use X as a predictor in our regression equation

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & if\ the\ ith\ sale\ was\ urban \\ \beta_0 + \epsilon_i & if\ the\ ith\ sale\ was\ rural \end{cases}$$

What does β_0 represent in the context of this problem? What does β_1 represent in the context of this problem?

Two Levels

A ***dummy variable*** is an indicator variable that takes on two possible numerical values.

Ex. For Urban, our dummy variable will be: *X* where

$$x_i = \begin{cases} 1 & if\ ith\ sale\ was\ urban \\ 0 & if\ the\ ith\ sale\ was\ rural \end{cases}$$

Now, we can use X as a predictor in our regression equation

$$y_i = \beta_0 + \beta_1 x_i + \in_i = \begin{cases} \beta_0 + \beta_1 + \in_i & if\ the\ ith\ sale\ was\ urban \\ \beta_0 + \in_i & if\ the\ ith\ sale\ was\ rural \end{cases}$$

What happens if I code the dummy variable 0/1 instead of 1/0? What if I code it 1/-1?

More than Two Levels

If a qualitative predictor has more than 2 levels, we can't use a single dummy variable to represent all possible values. But we can use multiple dummy variables.

More than Two Levels

If a qualitative predictor has more than 2 levels, we can't use a single dummy variable to represent all possible values. But we can use multiple dummy variables.

Ex. Let's say we want to use **`ShelveLoc`** (quality of shelving location at site <Good | Bad | Medium>) to predict `Sales`

More than Two Levels

If a qualitative predictor has more than 2 levels, we can't use a single dummy variable to represent all possible values. But we can use multiple dummy variables.

Ex. Let's say we want to use **`ShelveLoc`** (quality of shelving location at site <Good | Bad | Medium>) to predict `Sales`

$$x_{i1} = \begin{cases} 1 & if\ ith\ sale\ had\ good\ shelving \\ 0 & if\ ith\ sale\ did\ not\ have\ good\ shelving \end{cases}$$

$$x_{i2} = \begin{cases} 1 & if\ ith\ sale\ had\ bad\ shelving \\ 0 & if\ ith\ sale\ did\ not\ have\ bad\ shelving \end{cases}$$

More than Two Levels

If a qualitative predictor has more than 2 levels, we can't use a single dummy variable to represent all possible values. But we can use multiple dummy variables.

Ex. Let's say we want to use **`ShelveLoc`** (quality of shelving location at site <Good | Bad | Medium>) to predict `Sales`

$$x_{i1} = \begin{cases} 1 & \textit{if ith sale had good shelving} \\ 0 & \textit{if ith sale did not have good shelving} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \textit{if ith sale had bad shelving} \\ 0 & \textit{if ith sale did not have bad shelving} \end{cases}$$

Now,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

$$y_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \textit{if the ith sale had good shelving} \\ \beta_0 + \beta_2 + \epsilon_i & \textit{if the ith sale had bad shelving} \\ \beta_0 + \epsilon_i & \textit{if the ith sale had medium shelving} \end{cases}$$

What does each coefficient represent?

More than Two Levels

single dummy variable to represent all possible values. But we can use multiple dummy variables.

Ex. Let's say we want to use **`ShelveLoc`** (quality of shelving location at site <Good | Bad | Medium>) to predict `Sales`

$$x_{i1} = \begin{cases} 1 & if\ ith\ sale\ had\ good\ shelving \\ 0 & if\ ith\ sale\ did\ not\ have\ good\ shelving \end{cases}$$

$$x_{i2} = \begin{cases} 1 & if\ ith\ sale\ had\ bad\ shelving \\ 0 & if\ ith\ sale\ did\ not\ have\ bad\ shelving \end{cases}$$

Now,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

$$y_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & if\ the\ ith\ sale\ had\ good\ shelving \\ \beta_0 + \beta_2 + \epsilon_i & if\ the\ ith\ sale\ had\ bad\ shelving \\ \beta_0 + \epsilon_i & if\ the\ ith\ sale\ had\ medium\ shelving \end{cases}$$

More than Two Levels

Ex. Let's say we want to use **`ShelveLoc`** (quality of shelving location at site <Good | Bad | Medium>) to predict `Sales`

$$x_{i1} = \begin{cases} 1 & \textit{if ith sale had good shelving} \\ 0 & \textit{if ith sale did not have good shelving} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \textit{if ith sale had bad shelving} \\ 0 & \textit{if ith sale did not have bad shelving} \end{cases}$$

Now,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

$$y_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \textit{if the ith sale had good shelving} \\ \beta_0 + \beta_2 + \epsilon_i & \textit{if the ith sale had bad shelving} \\ \beta_0 + \epsilon_i & \textit{if the ith sale had medium shelving} \end{cases}$$

β_0 represents average sale for sales with medium shelving.

β_1 represents average difference in sale for sales with medium shelving vs good shelving .

β_2 represents average difference in sale for sales with medium shelving vs bad shelving .

Ex. Let's say we want to use **`ShelveLoc`** (quality of shelving location at site <Good | Bad | Medium>) to predict `Sales`

Notice we have 1 fewer dummy variables than levels.
The level with no dummy variable is known as the ***baseline***.

$$x_{i1} = \begin{cases} 1 & \text{if ith sale had good shelving} \\ 0 & \text{if ith sale did not have good shelving} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if ith sale had bad shelving} \\ 0 & \text{if ith sale did not have bad shelving} \end{cases}$$

More than Two Levels

Now,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

$$y_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if the ith sale had good shelving} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if the ith sale had bad shelving} \\ \beta_0 + \epsilon_i & \text{if the ith sale had medium shelving} \end{cases}$$

β_0 represents average sale for sales with medium shelving.

β_1 represents average difference in sale for sales with medium shelving vs good shelving .

β_2 represents average difference in sale for sales with medium shelving vs bad shelving .

Extensions of Linear Models

Assumptions

Standard linear regression is great for interpretable results, and works well on many real-world scenarios.

But it makes several assumptions that are often violated in practice.

Two important one are that the relationship between the predictors and response are

- *Additive*

and

- *Linear*

Additive Assumption

The ***additive assumption*** states that the association between a predictor, X_j, and the response Y does not depend on the values of the other predictors.

Additive Assumption

The ***additive assumption*** states that the association between a predictor, X_j, and the response Y does not depend on the values of the other predictors.

Remember our Advertising model

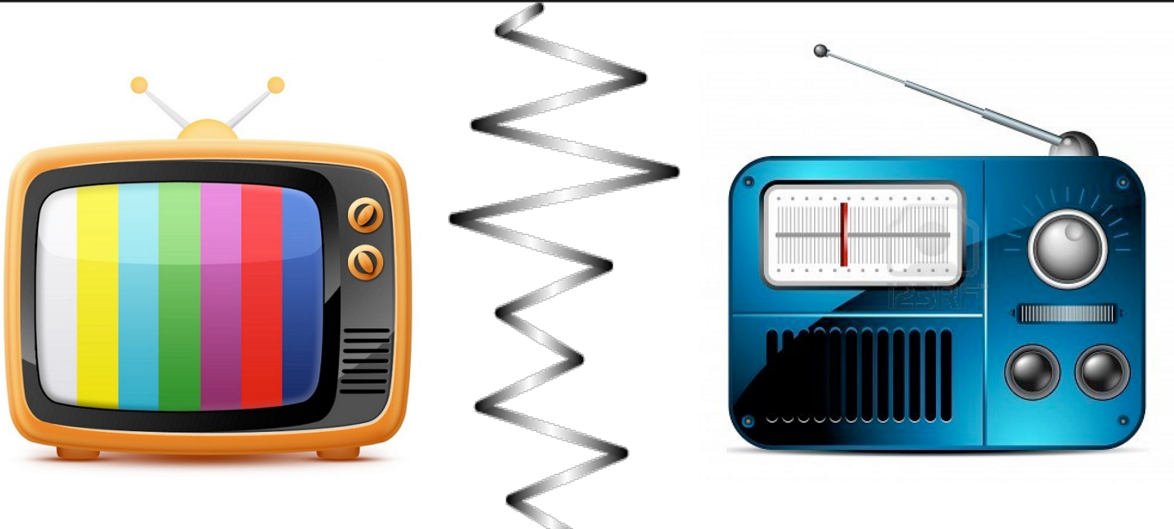
	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
~~newspaper~~	~~-0.001~~	~~0.0059~~	~~-0.18~~	~~0.8599~~

Additive Assumption

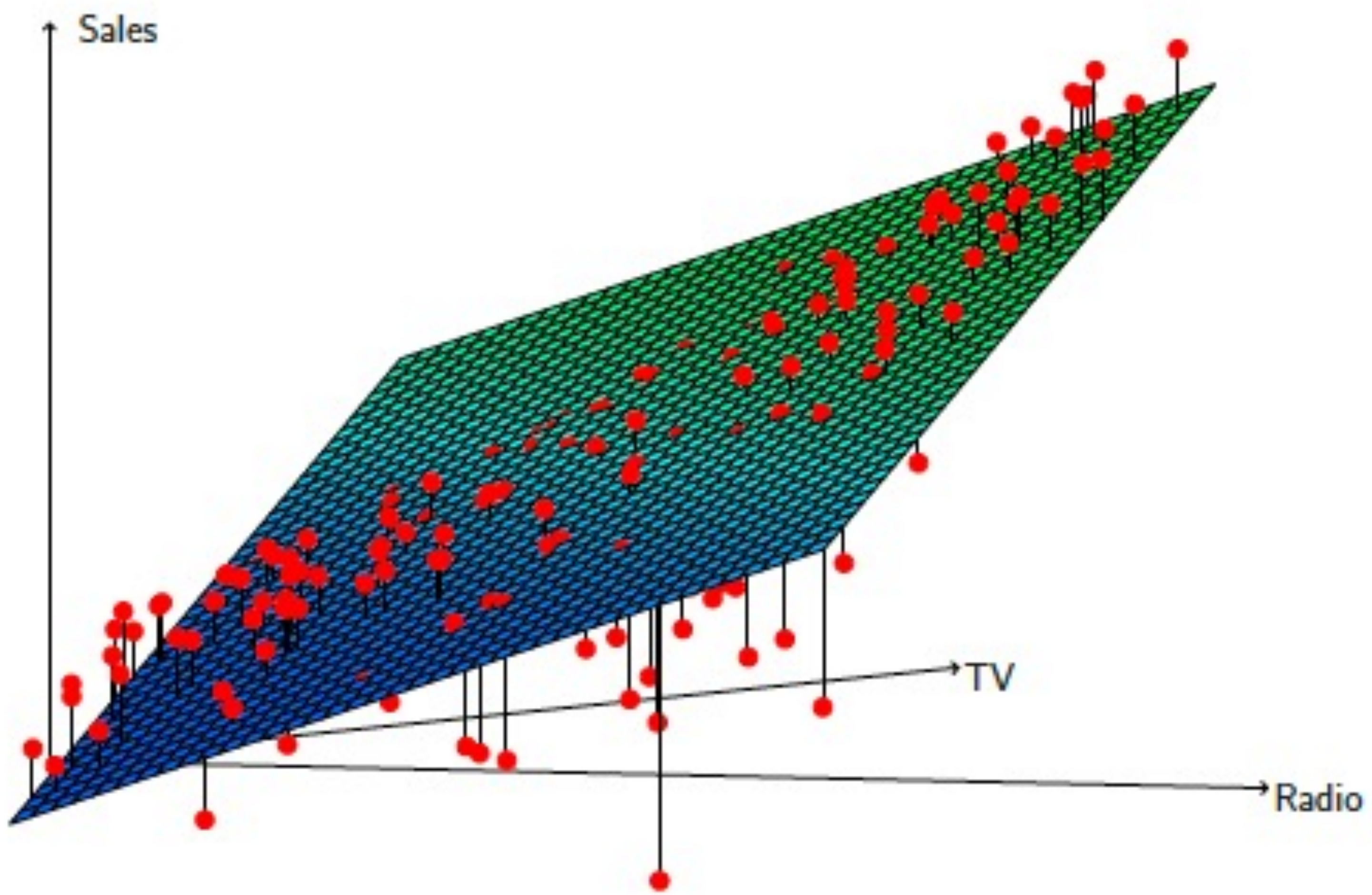
The ***additive assumption*** states that the association between a predictor, X_j, and the response Y does not depend on the values of the other predictors.

Remember our Advertising model

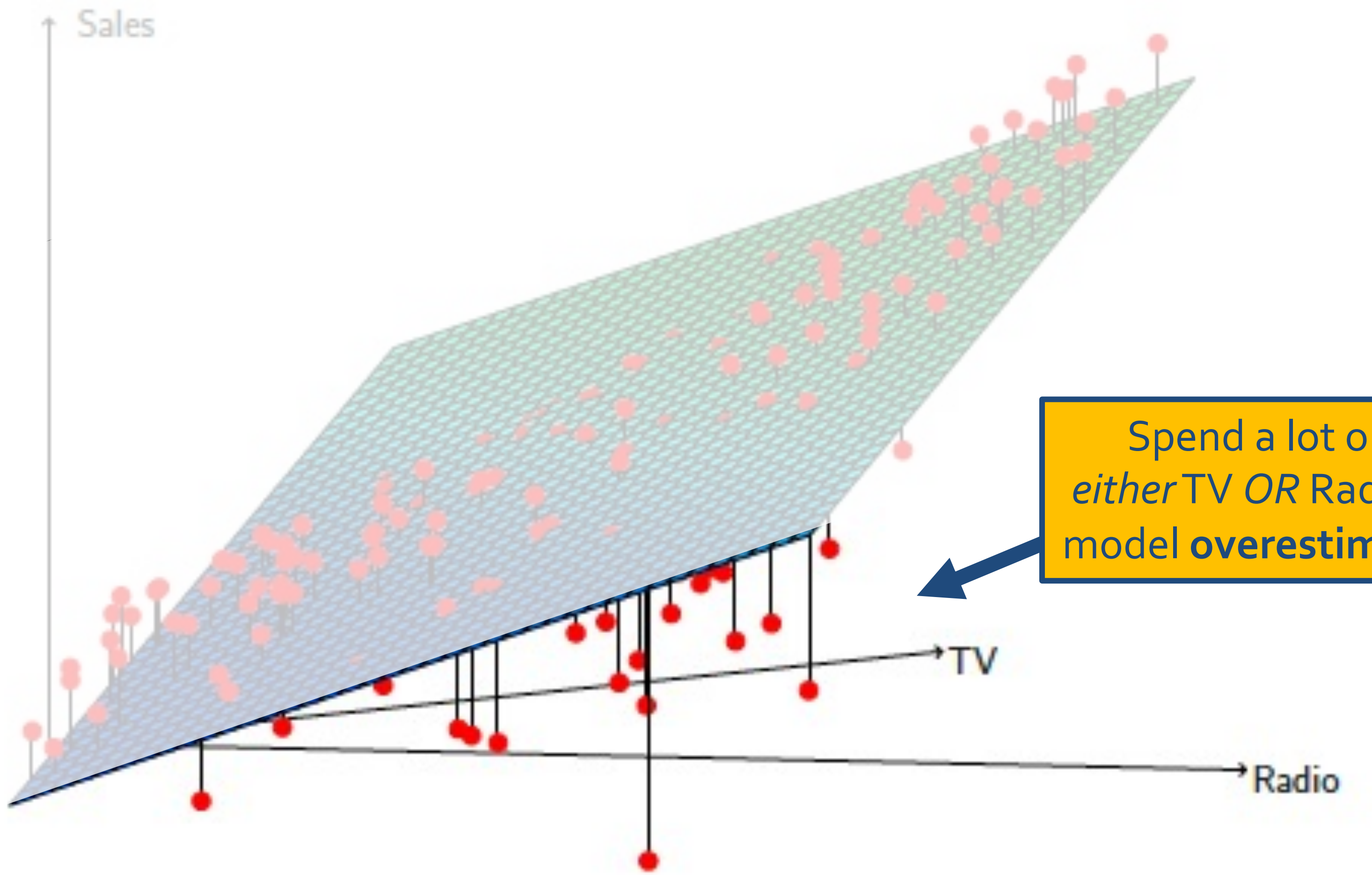
	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
~~newspaper~~	~~−0.001~~	~~0.0059~~	~~−0.18~~	~~0.8599~~



Interaction Effects

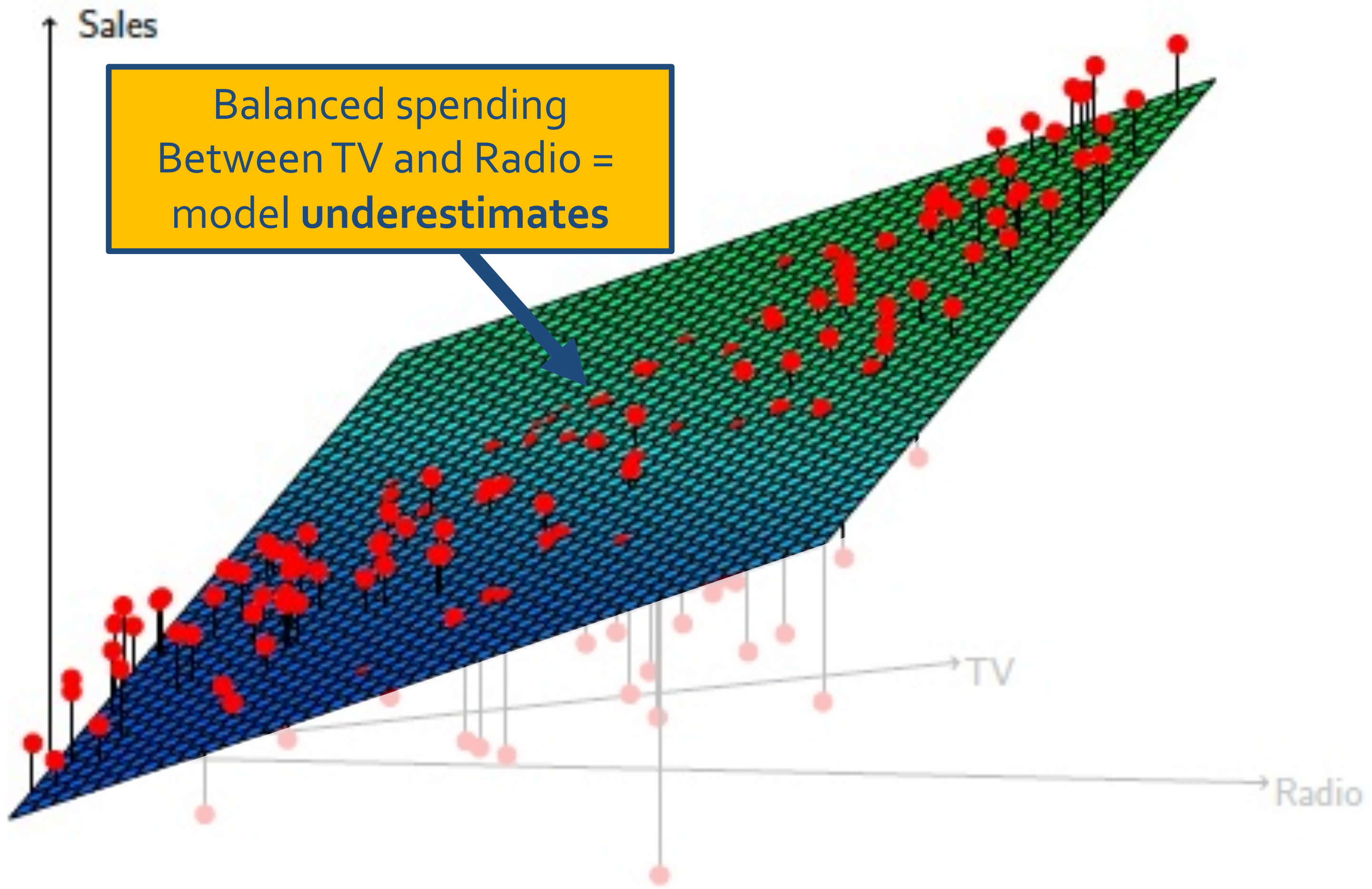


Interaction Effects



Spend a lot on *either* TV *OR* Radio = model **overestimates**

Interaction Effects



Interaction Effects

Our model does not account for the fact that spending money on `radio` advertising actually increases the effectiveness of TV advertising

$$Y = \beta_0 + \beta_1 \times radio + \beta_2 \times TV + \epsilon$$

But it can if we add an *interaction term*

$$Y = \beta_0 + \beta_1 \times radio + \beta_2 \times TV + \beta_3 \times (TV \times radio) + \epsilon$$

Interaction Effects

Our model does not account for the fact that spending money on `radio` advertising actually increases the effectiveness of TV advertising

$$Y = \beta_0 + \beta_1 \times radio + \beta_2 \times TV + \epsilon$$

But it can if we add an *interaction term*

$$Y = \beta_0 + \beta_1 \times radio + \beta_2 \times TV + \beta_3 \times (TV \times radio) + \epsilon$$

How does this fix the problem?
Hint: What is the new slope for radio?

Interaction Effects

Our model does not account for the fact that spending money on `radio` advertising actually increases the effectiveness of TV advertising

$$Y = \beta_0 + \beta_1 \times radio + \beta_2 \times TV + \epsilon$$

But it can if we add an *interaction term*

$$Y = \beta_0 + \beta_1 \times radio + \beta_2 \times TV + \beta_3 \times (TV \times radio) + \epsilon$$
$$Y = \beta_0 + (\beta_1 + \beta_3 \times TV) \times radio + \beta_2 \times TV + \epsilon$$
$$Y = \beta_0 + \widetilde{\beta_1} \times radio + \beta_2 \times TV + \epsilon$$

Interaction Effects

If we fit this model

$$Y = \beta_0 + \beta_1 \times radio + \beta_2 \times TV + \beta_3 \times (TV \times radio) + \epsilon$$

$$Y = \beta_0 + (\beta_1 + \beta_3 \times TV) \times radio + \beta_2 \times TV + \epsilon$$

$$Y = \beta_0 + \widetilde{\beta_1} \times radio + \beta_2 \times TV + \epsilon$$

	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

Interaction Effects

If we fit this model

$$Y = \beta_0 + \beta_1 \times radio + \beta_2 \times TV + \beta_3 \times (TV \times radio) + \epsilon$$

$$Y = \beta_0 + (\beta_1 + \beta_3 \times TV) \times radio + \beta_2 \times TV + \epsilon$$

$$Y = \beta_0 + \widetilde{\beta_1} \times radio + \beta_2 \times TV + \epsilon$$

	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

What does the small p-value for TV×radio indicate?

Interaction Effects

If we fit this model

$$Y = \beta_0 + \beta_1 \times radio + \beta_2 \times TV + \beta_3 \times (TV \times radio) + \epsilon$$

$$Y = \beta_0 + (\beta_1 + \beta_3 \times TV) \times radio + \beta_2 \times TV + \epsilon$$

$$Y = \beta_0 + \widetilde{\beta_1} \times radio + \beta_2 \times TV + \epsilon$$

	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

R^2 without interaction term is 89.7%; this model: 96.8%

Interaction Effects

If we fit this model

$$Y = \beta_0 + \beta_1 \times radio + \beta_2 \times TV + \beta_3 \times (TV \times radio) + \epsilon$$

$$Y = \beta_0 + (\beta_1 + \beta_3 \times TV) \times radio + \beta_2 \times TV + \epsilon$$

$$Y = \beta_0 + \widetilde{\beta_1} \times radio + \beta_2 \times TV + \epsilon$$

	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

R^2 without interaction term is 89.7%; this model: 96.8%

diff. var. explained by each model

var. missed by first model

$$\frac{96.8 - 89.7}{100 - 89.7} = 69\%$$

of the variability that our previous model missed is explained by the interaction term.

Interaction Effects

If we fit this model

$$Y = \beta_0 + \beta_1 \times radio + \beta_2 \times TV + \beta_3 \times (TV \times radio) + \epsilon$$
$$Y = \beta_0 + (\beta_1 + \beta_3 \times TV) \times radio + \beta_2 \times TV + \epsilon$$
$$Y = \beta_0 + \widetilde{\beta_1} \times radio + \beta_2 \times TV + \epsilon$$

	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

- In this example, p-values for all predictors are significant
- This doesn't always happen
- ***Hierarchical principle***: if we include and interaction term, we should include the main effects too (regardless of significance)

Linearity Assumption

The ***linearity assumption*** states that the change in the response, Y, associated with a one-unit change in X_j is constant, regardless of the value of X_j.

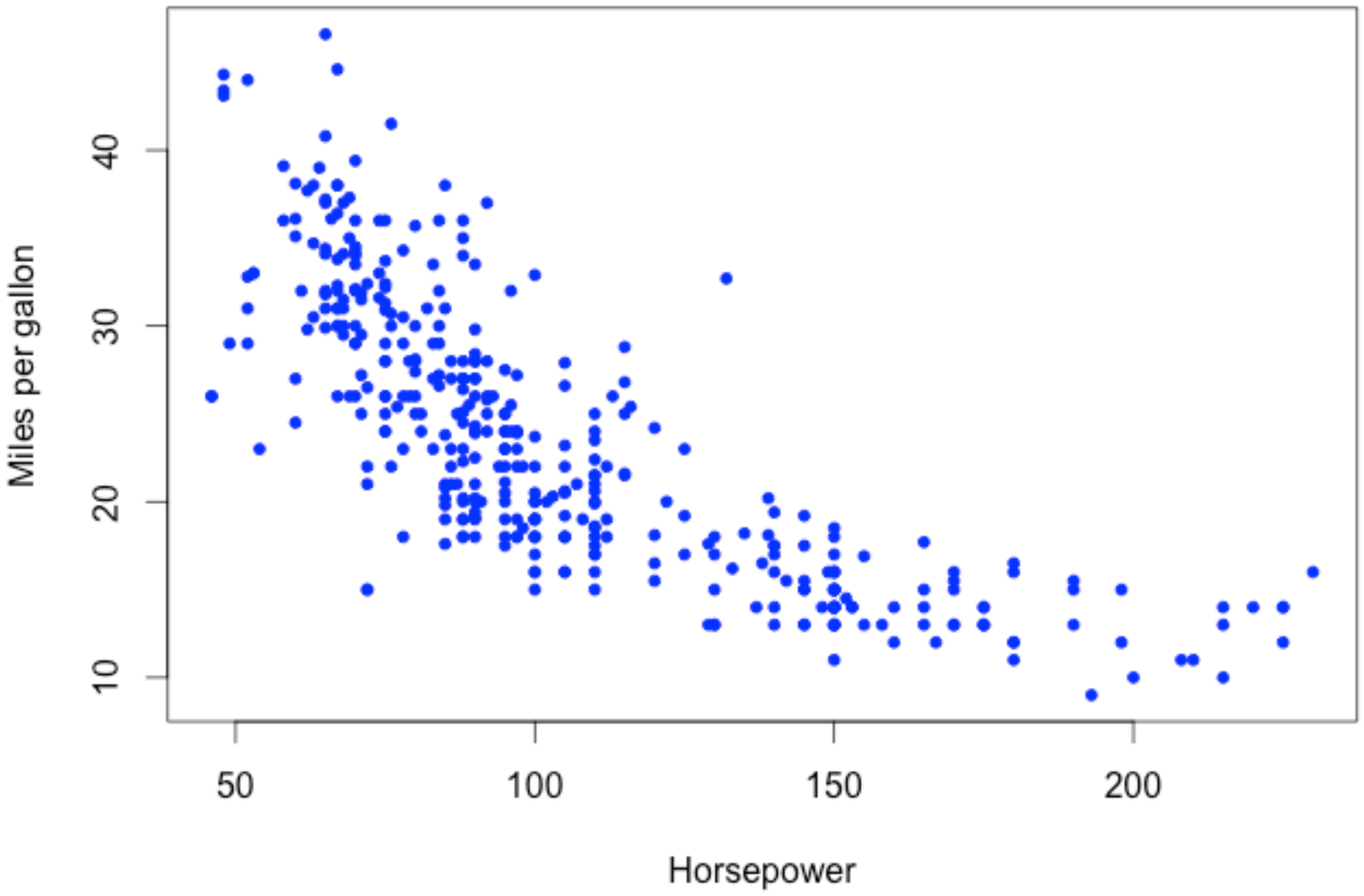
Linearity Assumption

The ***linearity assumption*** states that the change in the response, Y, associated with a one-unit change in X_j is constant, regardless of the value of X_j.

- If the true relationship is far from linear:
  - The conclusions we draw from a linear model are probably flawed
  - The prediction accuracy of the model is likely pretty low

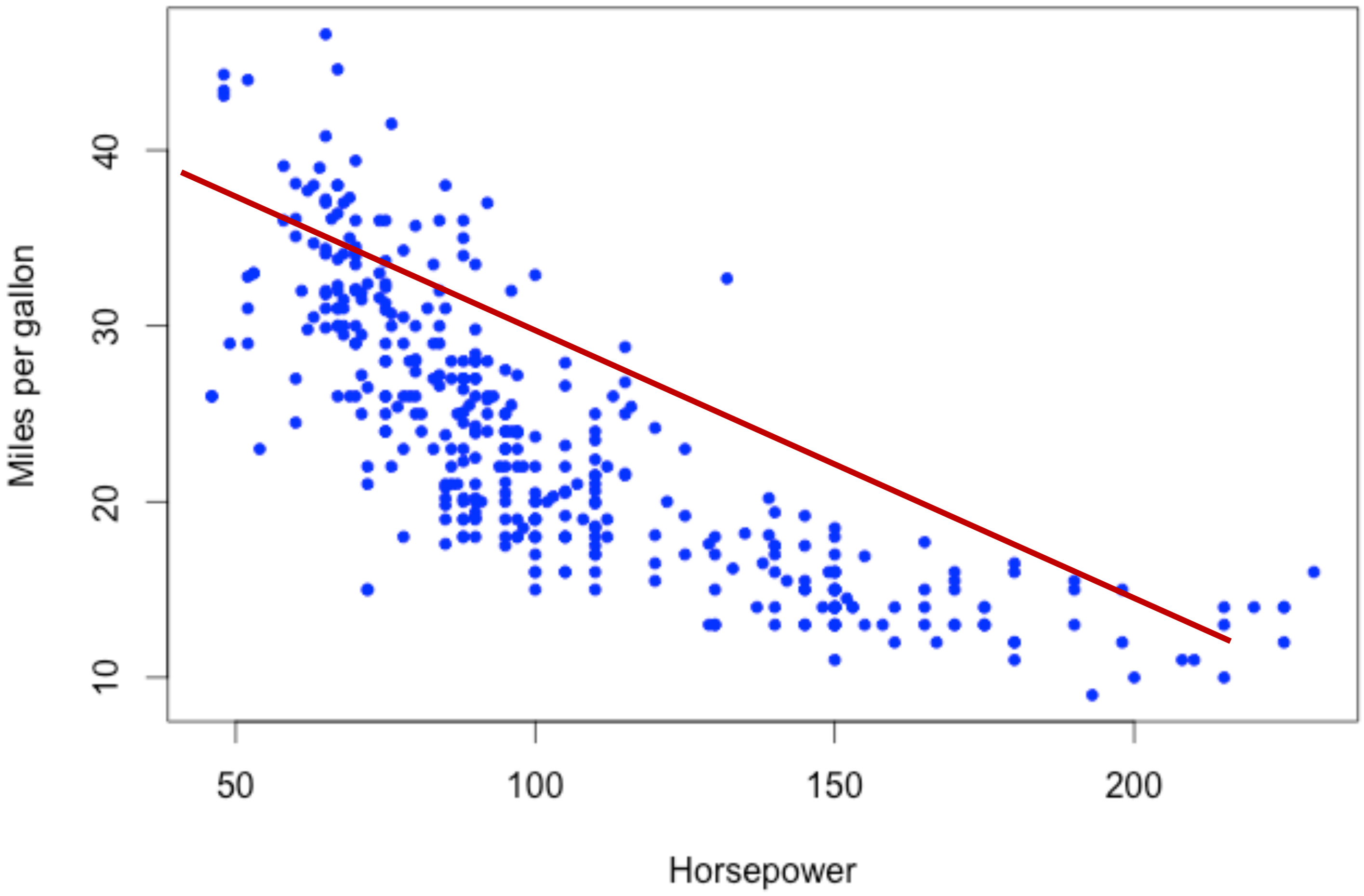
Linearity Assumption

Example:



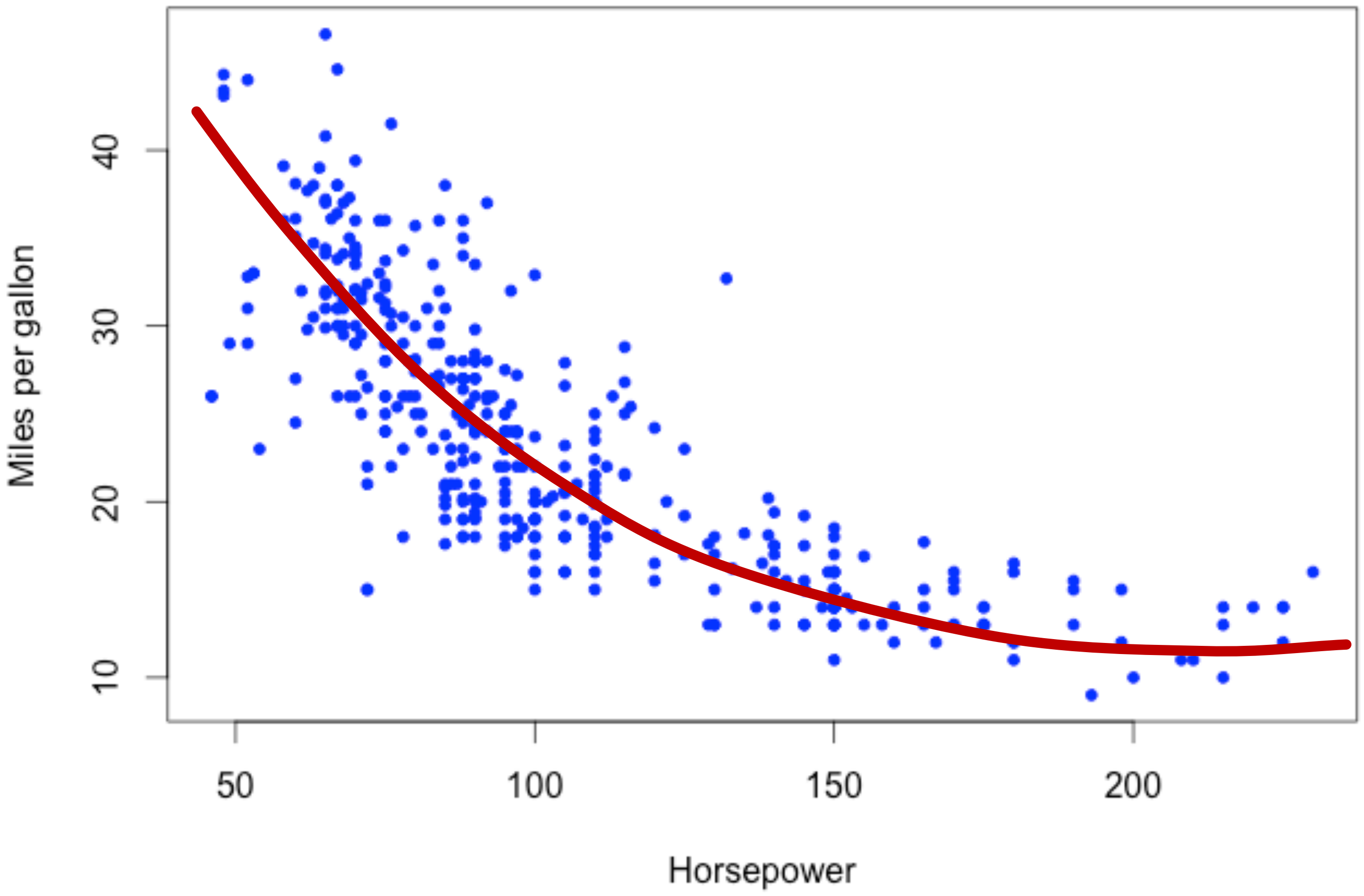
Linearity Assumption

Example:



Linearity Assumption

Example:



Polynomial Regression

- **Simple approach**: use polynomial transformations

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

- **Note**: still a linear model! (X_2 = horesepower^2)

Polynomial Regression

- **Simple approach**: use polynomial transformations

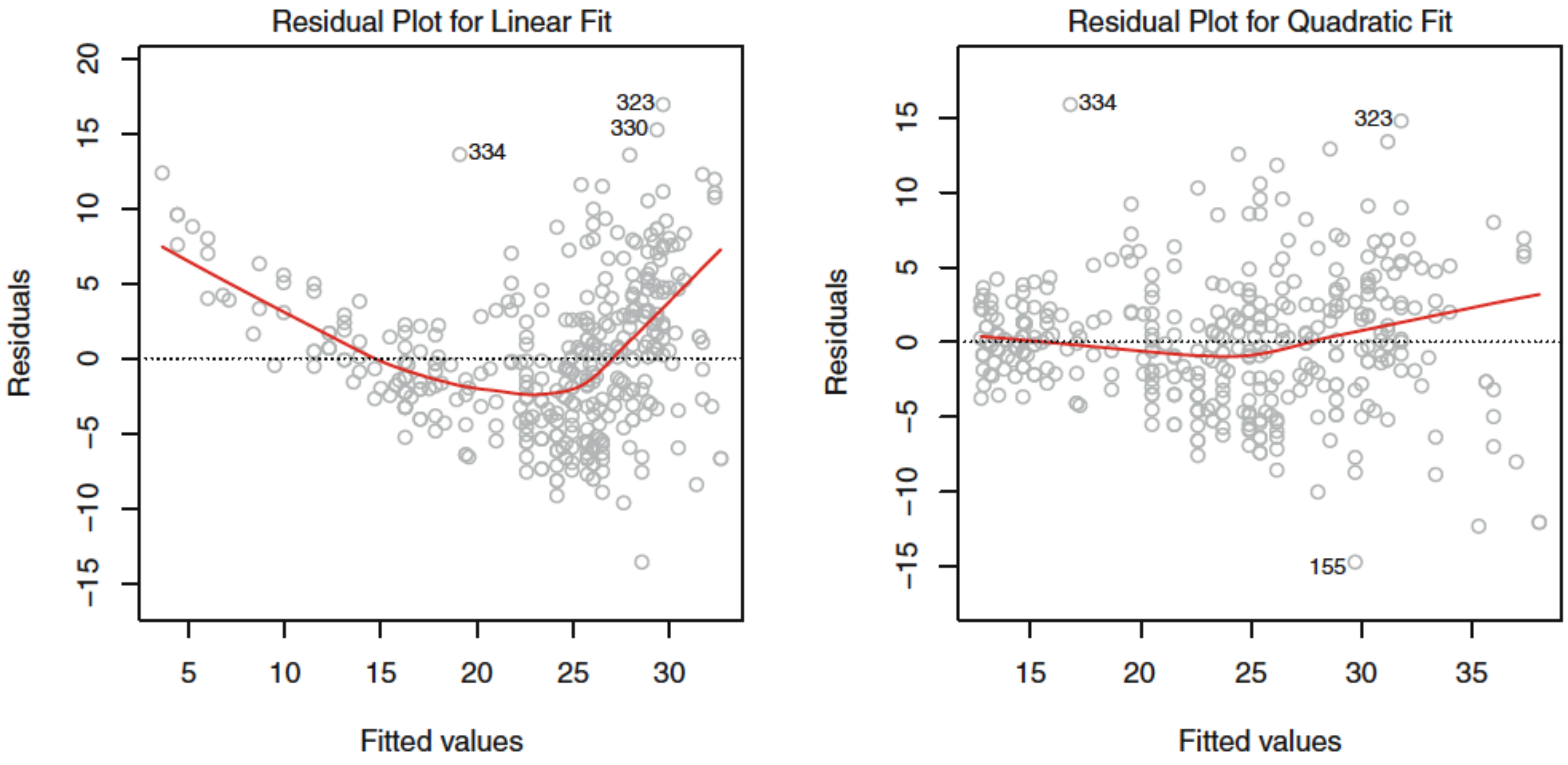
$$\texttt{\textbf{mpg}} = \beta_0 + \beta_1 \times \texttt{\textbf{horsepower}} + \beta_2 \times \texttt{\textbf{horsepower}}^2 + \epsilon$$

- **Note**: still a linear model! (X_2 = horesepower^2)

How do we know the correct power?!

How to tell if you need more power

- Residual plots can help identify problem areas in the model by highlighting patterns in errors



Potential Problems

Breaking Linear Regression

- Potential issues
  1. Correlated error terms
  2. Non-constant variance of error terms
  3. Outliers
  4. High leverage points
  5. Collinearity

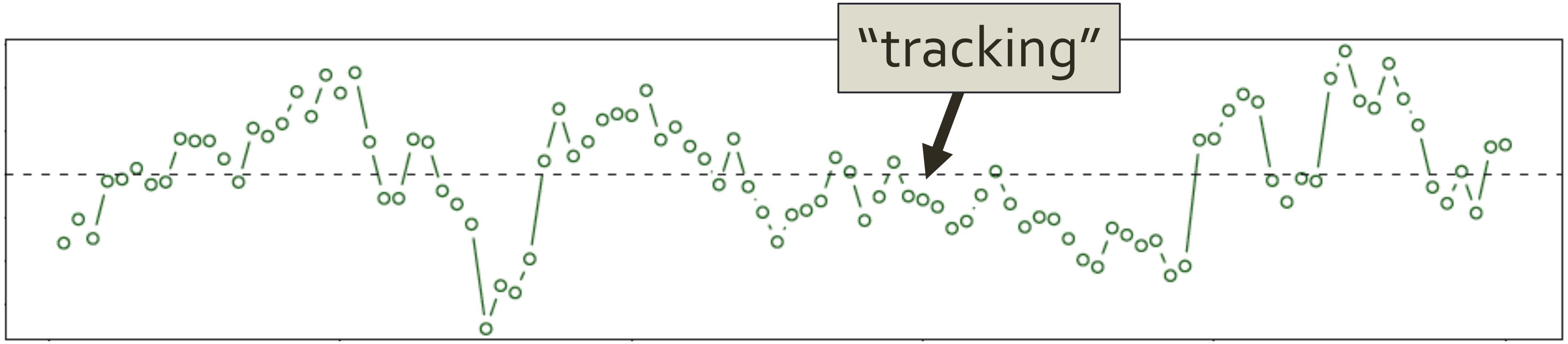
Correlated Error Terms

- LR assumes that the error terms are **uncorrelated**
- If these terms *are* correlated, the estimated standard error will tend to **underestimate** the true standard error. As a result,
  - CI's will be narrower than they should be and
  - p-values will be lower than they should be

Correlated Error Terms

Checking for correlated error terms in time-series data

- Plot residuals as a function of time
- If error are uncorrelated, there will be no discernable pattern
- If errors are correlated, we will see *tracking* (adjacent residuals with similar values)



Non-constant variance of error terms

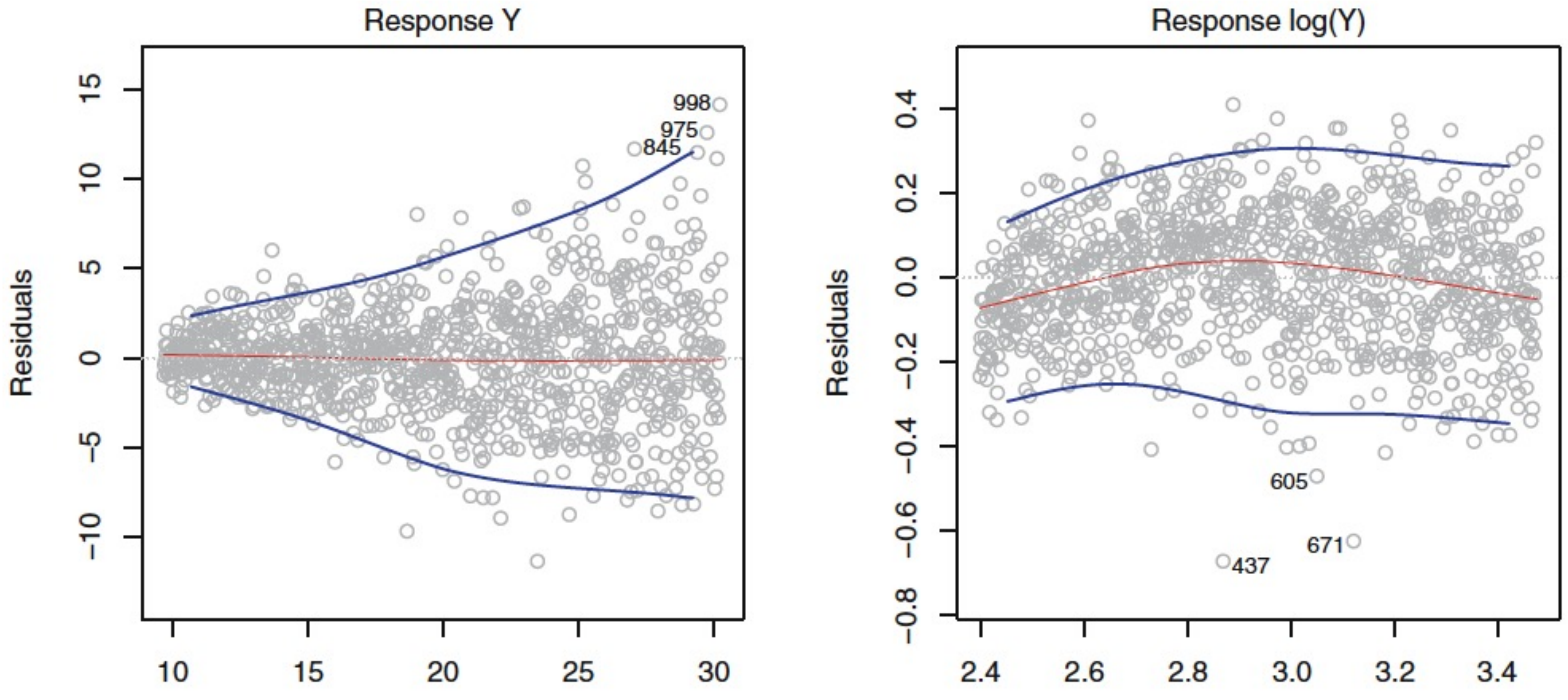
- LR assumes that error terms have constant variance:

$$Var(e_i) = \sigma^2$$

- SE's, CI's, and hypothesis tests rely on this assumption
- Often not the case (e.g. error terms might increase with the value of the response)
- Non-constant variance in errors is called ***heteroscedasticity***

Identify and Fix Heteroscedasticy

- Identifying: The residuals plot will show a funnel shape



- Fixing:
  - transform the response using a **concave function** (like *log* or *sqrt*)
  - **weight** the observations proportional to the inverse variance

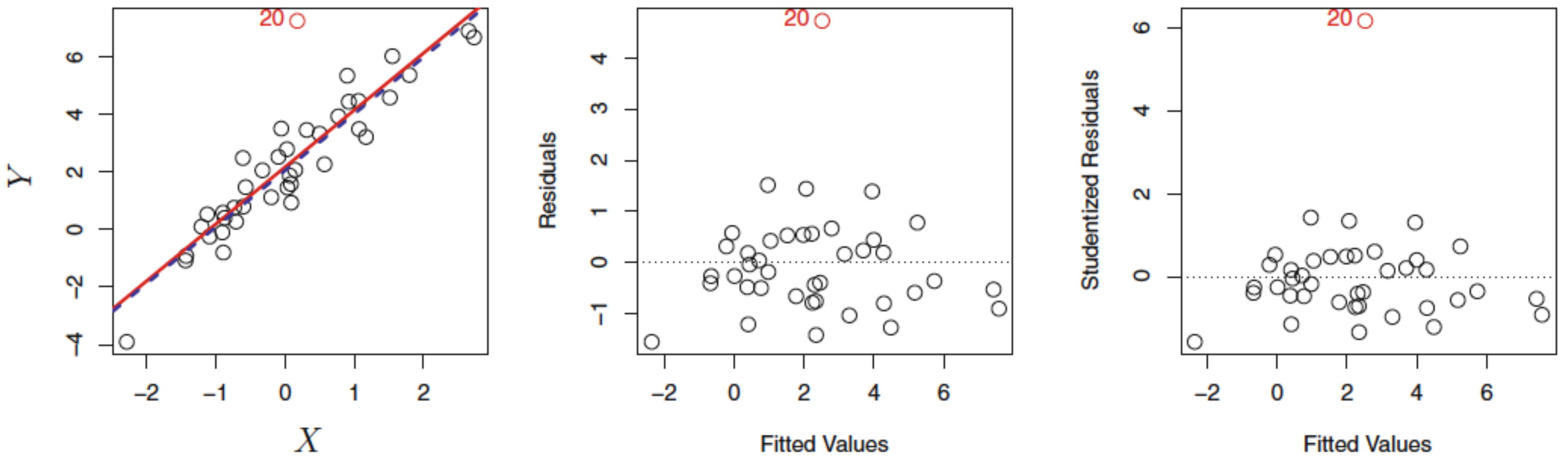
Outliers

An ***outlier*** is an observation whose true response is REALLY FAR from the one predicted by the model

- Sometimes indicate a problem with the model (i.e. a missing predictor), or might just be a data collection error
- Can mess with RSE and R^2, which can lead us to misinterpret the model's fit

Identify and Fix Outliers

Identify: Residual plots can help identify outliers, but sometimes it's hard to pick a cutoff point (how far is "too far"?)



Fix:

- Divide each residual by dividing by its estimated standard error (*studentized residuals*), and flag anything larger than 3 in absolute value

High Leverage Points

- Outliers are unusual values in the response
- ***High leverage points are*** unusual values in the predictor(s)

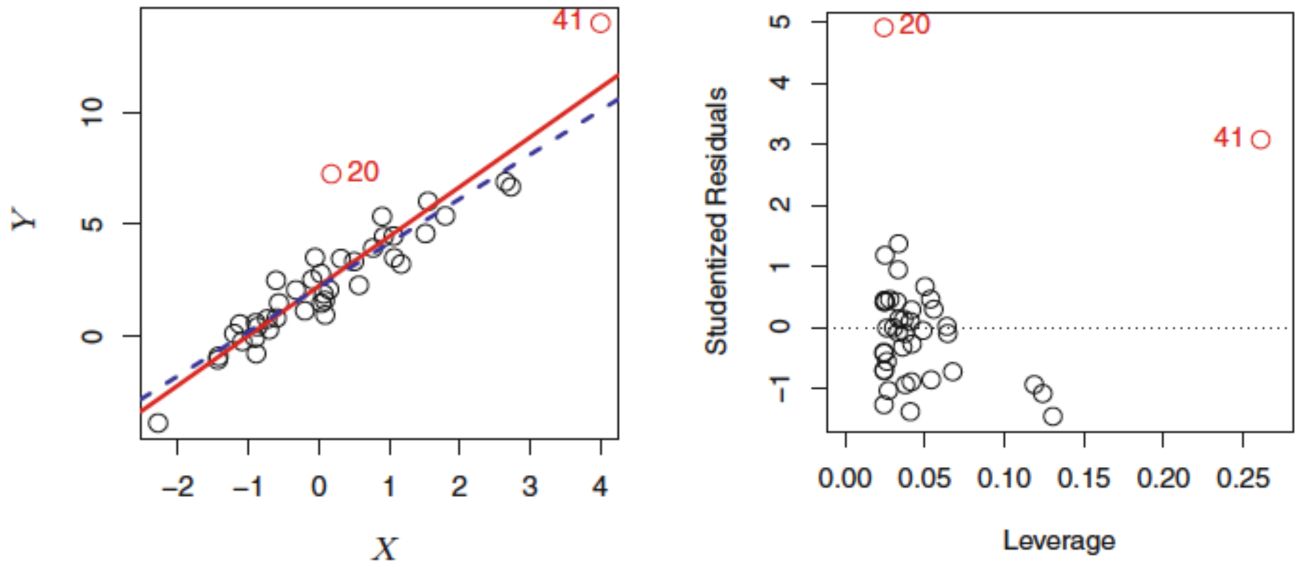
  - The more predictors you have, the harder they can be to spot
  - These points can have a major impact on the least squares line, which could invalidate the entire fit
    - We don't want only one or a few inputs to cause large changes in the entire model

Identify High Leverage Points

Compute the leverage statistic. For SLR:

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

- The leverage statistic is always a value between $\frac{1}{n}$ and n
- The average for all observations is $\frac{p+1}{n}$, if a statistic is much greater than the average, the point is probably a high leverage point



Collinearity

Collinearity is when two or more predictor variables are closely related to one another

- This makes it hard to isolate the individual effects of each predictor, which increases uncertainty in coefficient estimates

- As a result, it is harder to detect whether or not an effect is actually present (because SE has increased)

Identify Collinearity

- Look at the correlation matrix of the predictors
- **`Auto`** dataset: just about everything is highly correlated

	mpg	cylinders	displacement	horsepower	weight	acceleration	year
mpg	1	-0.7776175	-0.8051269	-0.7784268	-0.8322442	0.4233285	0.5805410
cylinders		1	0.9508233	0.8429834	0.8975273	-0.5046834	-0.3456474
displacement			1	0.8972570	0.9329944	-0.5438005	-0.3698552
horsepower				1	0.8645377	-0.6891955	-0.4163615
weight					1	-0.4168392	-0.3091199
acceleration						1	0.2903161
year							1
origin							

- **Note**: ***multicollinearity*** is when more than two variables are correlated; this will not show in this correlation matrix.

Dealing with Collinearity

Options include:

- Drop one of the problematic variables from the regression (collinearity implies they're redundant)
- Combine collinear variables into a single predictor (ex. take the average)