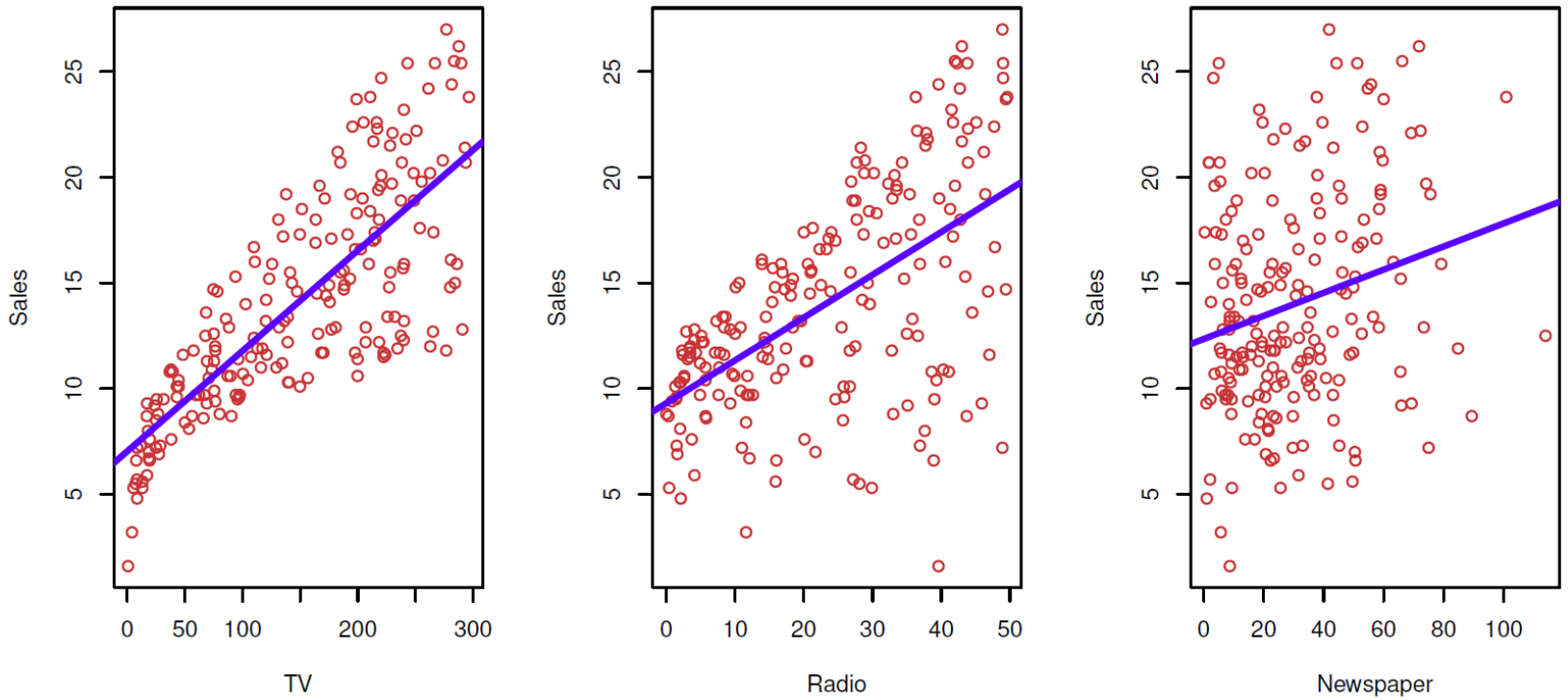
LINEAR REGRESSION

ADVERTISING DATA



$$Sales \approx f(TV, Radio, Newspaper)$$

MAIN QUESTIONS FOR A MARKETING PLAN

1. **Is there a relationship between advertising budget and sales?**
   - If evidence is weak, then one might argue that no money should be spent on advertising
2. **How strong is the relationship?**
   - Can we predict sales with a high level of accuracy?
3. **Which media contribute to sales?**
   - Do all three media, or do just one or two of the media?
4. **How accurately can we estimate the effect of each medium on sales?**
   - For every dollar spent on advertising in a particular medium, by what amount will sales increase? How accurately can we predict this amount of increase

MAIN QUESTIONS FOR A MARKETING PLAN

5. **How accurately can we predict future sales?**
   - For any given level of TV, Radio, or Newspaper advertising, what is our prediction for sales, and what is the accuracy of this prediction?
6. **Is the relationship linear?**
7. **Is there synergy among the advertising media?**
   - Perhaps spending $50,000 on TV and $50,000 on Radio results in more sales than allocating $100,000 to either TV or Radio individually.
   - In marketing, this is known as synergy effect
   - In statistics it is called an interaction effect

SIMPLE LINEAR REGRESSION

SIMPLE LINEAR REGRESSION

- Predicting quantitative response Y by a single predictor X

$$Y = \beta_0 + \beta_1 X + e$$

- If $X = TV, Y = sales$

$$sales = \beta_0 + \beta_1 TV + e$$

- β_0 -intercept, β_1- slope

SIMPLE LINEAR REGRESSION

- Prediction model

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- Estimating intercept and slope $(\hat{\beta}_0, \hat{\beta}_1)$ based on training data $(x_1, y_1), \dots, (x_n, y_n)$
- The least squares criterion

$$I(\hat{\beta}_0, \hat{\beta}_1) = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 = \sum_{k=1}^{n} (y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k)^2 \rightarrow min$$

- If $e_k = \hat{y}_k - y_k$

$$Residual\ sum\ of\ squares = RSS = \sum_{k=1}^{n} e_k^2 = e_1^2 + \cdots + e_n^2 \rightarrow min$$

ESTIMATION OF THE COEFFICIENTS

$$\frac{\partial I}{\partial \hat{\beta}_0} = -2\sum_{k=1}^{n}\left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right) = 0$$

$$\frac{\partial I}{\partial \hat{\beta}_1} = -2\sum_{k=1}^{n}\left(y_k - \hat{\beta}_0 - \hat{\beta}_1 x_k\right)x_k = 0$$

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

$$\hat{\beta}_1 = \frac{\sum_{k=1}^{n}(x-\bar{x})(y-\bar{y})}{\sum_{k=1}^{n}(x-\bar{x})^2} = \frac{cov(X,Y)}{\sigma_x \sigma_y}\frac{\sigma_y}{\sigma_x} = r\frac{\sigma_y}{\sigma_x}$$

GRADIENT VECTOR AND HESSIAN MATRIX

$$y = f(x_1, \dots, x_p)$$

$$grad(f) = \nabla f = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_p}\right)$$

$$H = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_p \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_p^2} \end{pmatrix}$$

$$H_{ij}(x_1, \dots, x_p) = \frac{\partial^2 f}{\partial x_i \partial x_j}, \qquad i, j = 1, \dots, p$$

NECESSARY AND SUFFICIENT CONDITIONS FOR A FUNCTION TO HAVE AN EXTREMUM

- Necessary condition (stationary point):

$$\nabla f = 0 \; at \; x = x^0 = (x_1^0, \dots, x_p^0)$$

- Sufficient condition (point of extrema):

If the quadratic form

$$Q(x_1, \dots, x_p) = \sum_{i=1}^{p} \sum_{j=1}^{p} H_{ij}(x^0) x_i x_j$$

is positive (negative) definite - takes only positive values – then $x = x^0$ is the point of minimum (maximum)

SYLVESTER CONDITION

- Necessary and sufficient conditions for a quadratic form

$$Q(x_1, \dots, x_p) = \sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij} x_i x_j \,,$$

to be positive definite are

$$a_{11} > 0, \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} > 0, \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} > 0, \dots, \begin{vmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{p1} & \dots & a_{pp} \end{vmatrix} > 0$$

WHY WE GET MINIMUM AT STATIONARY POINTS?

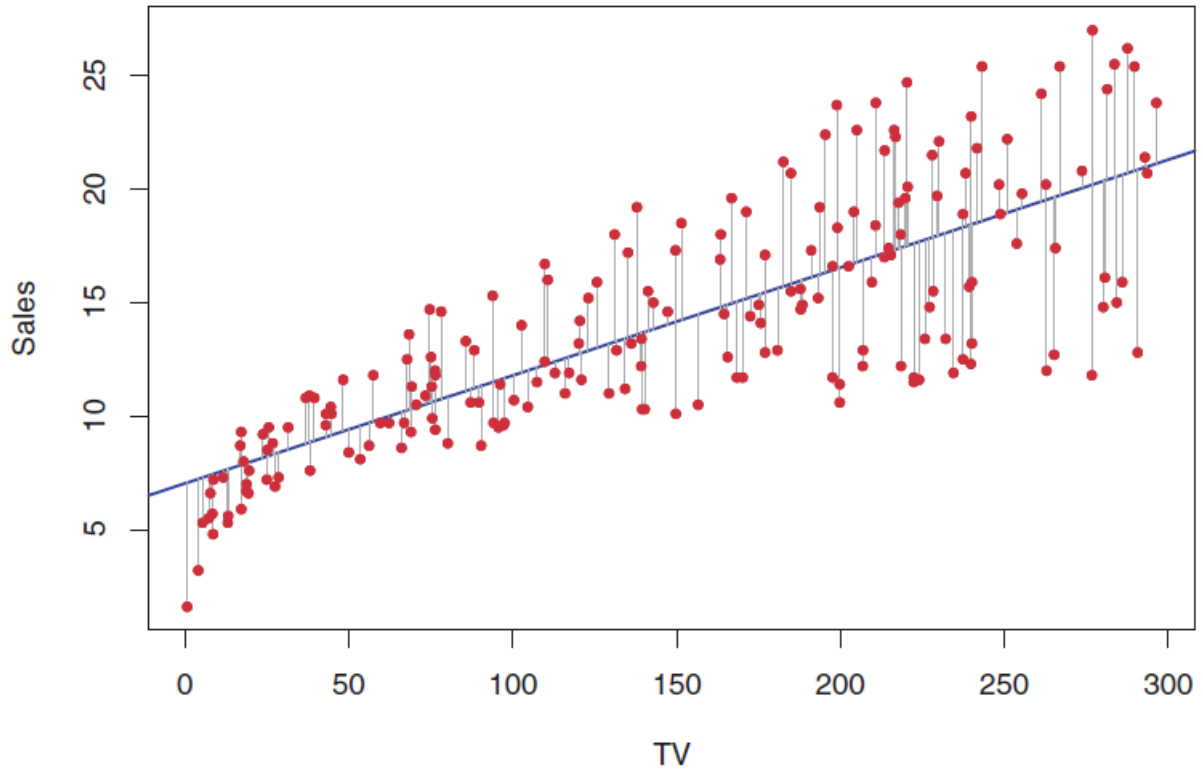
$$\frac{\partial^2 I}{\partial \hat{\beta}_0^2} = 2n > 0,$$

$$\frac{\partial^2 I}{\partial \hat{\beta}_1^2} = 2\sum_{k=1}^{n} x_k^2 ,$$

$$\frac{\partial^2 I}{\partial \hat{\beta}_0 \partial \hat{\beta}_1} = 2\sum_{k=1}^{n} x_k$$

$$\frac{\partial^2 I}{\partial \hat{\beta}_0^2}\frac{\partial^2 I}{\partial \hat{\beta}_1^2} - \left(\frac{\partial^2 I}{\partial \hat{\beta}_0 \partial \hat{\beta}_1}\right)^2 = 4n^2\left(\frac{1}{n}\sum_{k=1}^{n} x_k^2 - \left(\frac{1}{n}\sum_{k=1}^{n} x_k\right)^2\right) > 0$$

SALES ~ TV

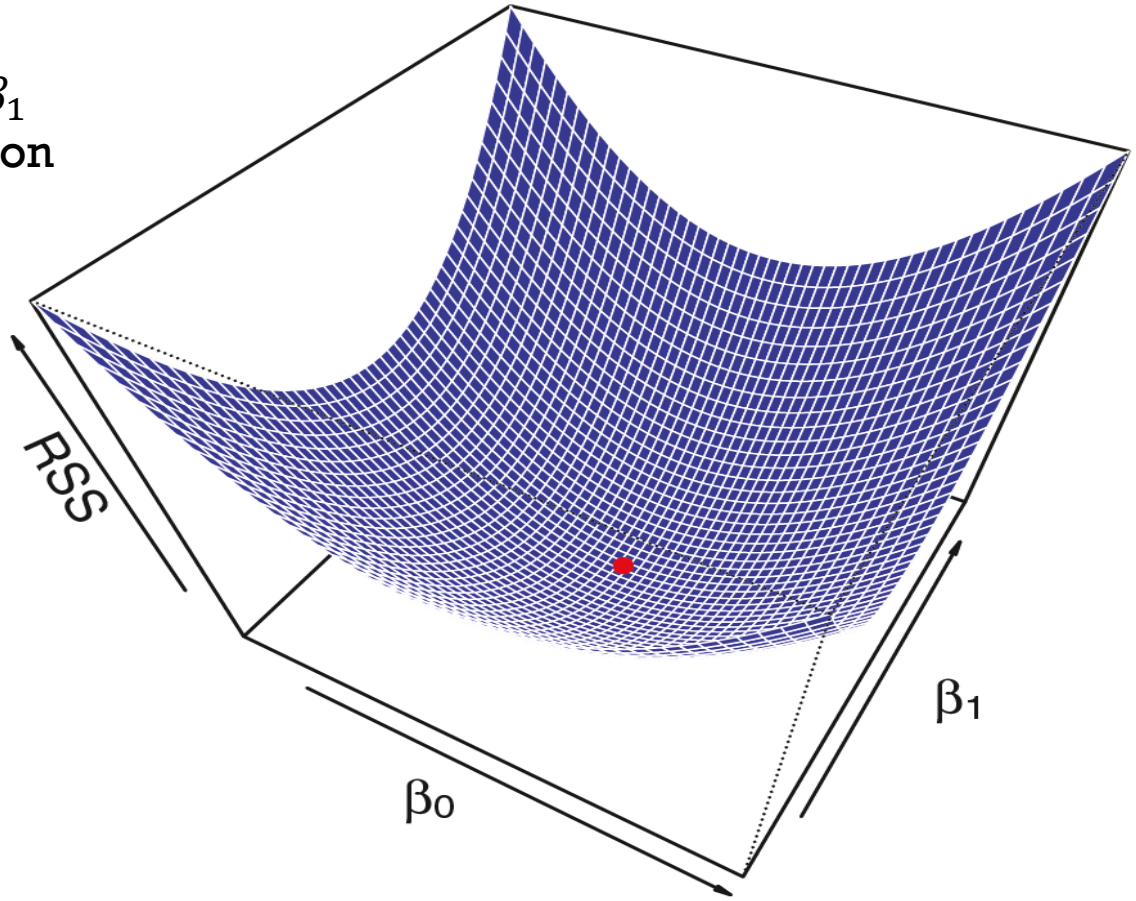


$\hat{\beta}_0 = 7.03, \quad \hat{\beta}_1 = 0.0475, \quad sales = 7.03 + 0.0475\, TV$

Additional $1000 invested for TV advertising will help to sell additional 47.5 units of the product

RSS FOR ADVERTISING DATA

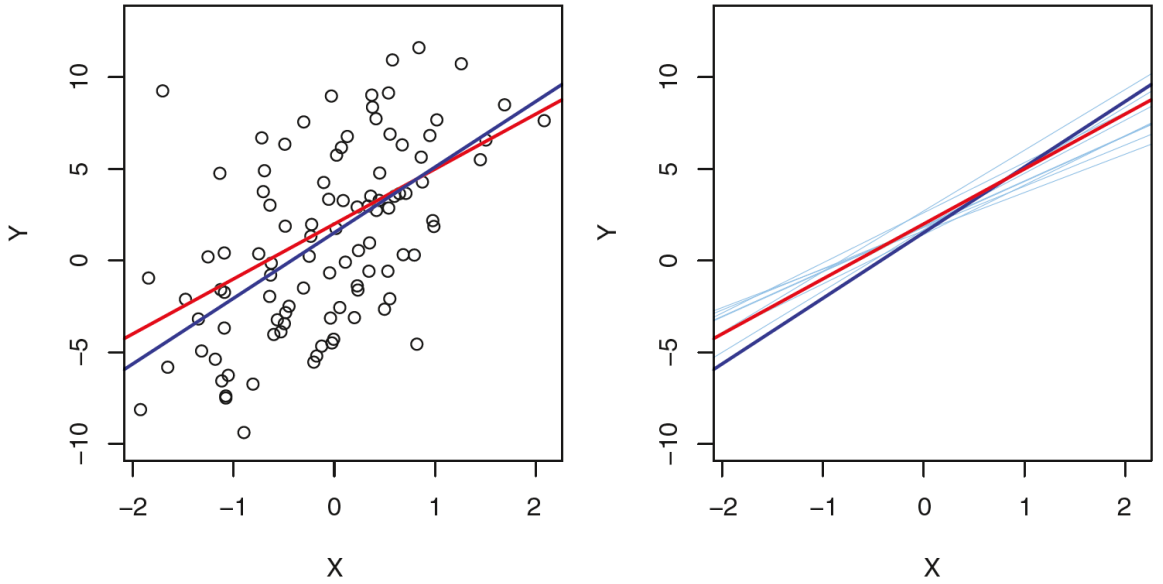
Red dot corresponds to β_0 and β_1 derived from the linear regression



ACCURACY OF THE COEFFICIENTS ESTIMATES

$Y = \beta_0 + \beta_1 X$ – population regression line

$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ – least squares line



Red line: true relationship $Y = 2 + 3X$
Black points: observed data
Blue line: regression by an observed data $Y = 2 + 3X + e$

STANDARD ERRORS OF THE COEFFICIENTS

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

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

$$\sigma^2 = var(e)$$

- Estimate of σ is known as the residual standard error

$$RSE = \sqrt{RSS/(n-2)}$$

ACCURACY OF THE COEFFICIENTS ESTIMATES

- Confidence intervals for the coefficients

$$\beta_0 \in \left[\hat{\beta}_0 - z \cdot SE\left(\hat{\beta}_0\right), \hat{\beta}_0 + z \cdot SE\left(\hat{\beta}_0\right)\right],$$

$$\beta_1 \in \left[\hat{\beta}_1 - z \cdot SE\left(\hat{\beta}_1\right), \hat{\beta}_1 + z \cdot SE\left(\hat{\beta}_1\right)\right]$$

- $z = 2$ for 95% confidence interval

ADVERTISING DATA

- For sales ~ TV model

$$SE(\hat{\beta}_0) = 0.9, SE(\hat{\beta}_1) = 0.0055$$

$$\beta_0 \in [6.130, 7.935], \beta_1 \in [0.042, 0.053]$$

- We can conclude that in the absence of any advertising, sales will, on average, fall somewhere between 6.130 and 7.940 units
- Each $\$1{,}000$ increase in TV advertising will lead to an average increase in sales of between 42 and 53 units

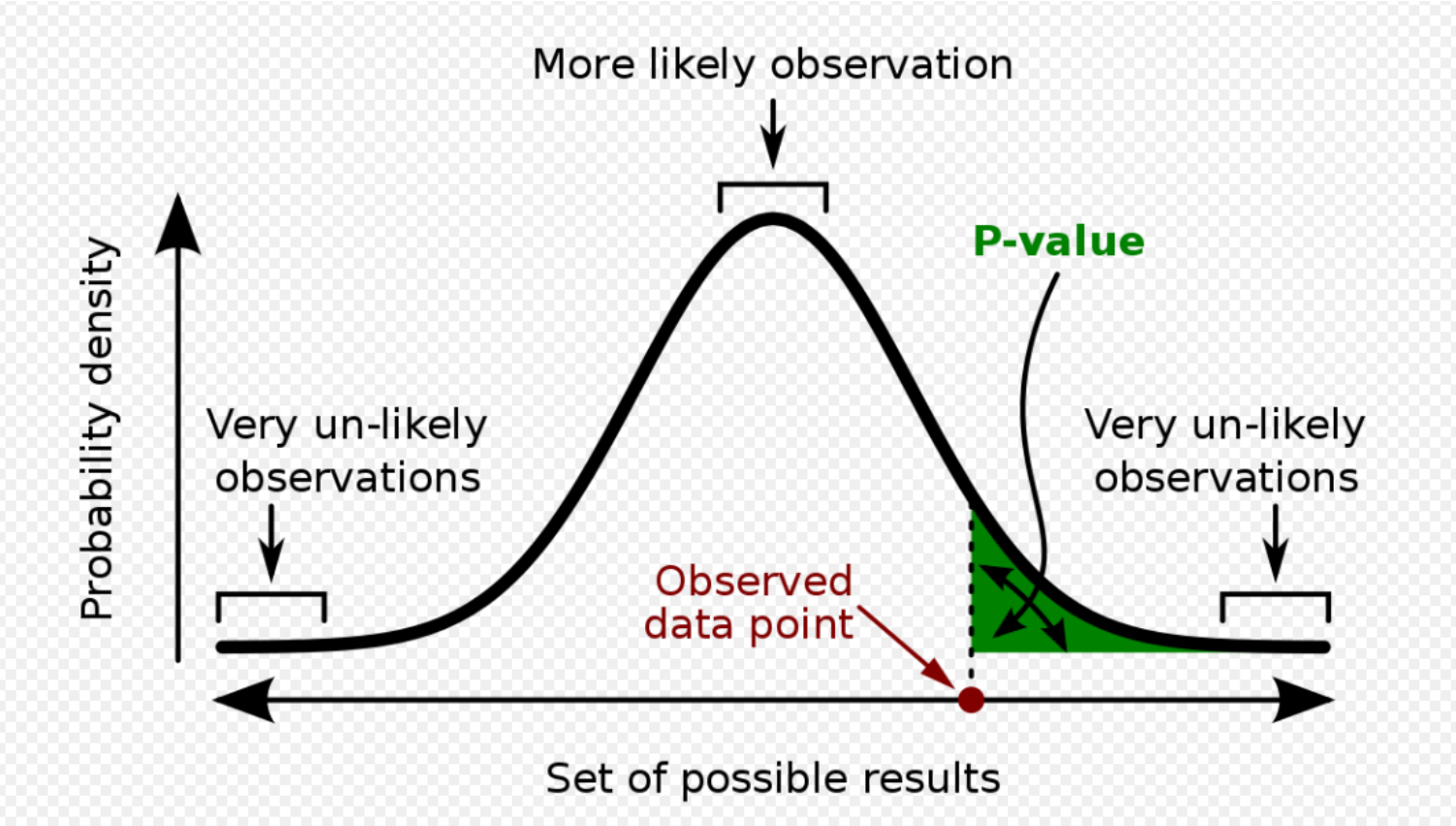
IS THERE ANY RELATIONSHIP BETWEEN X AND Y?

- Null Hypothesis: H_0: There is no relationship ($\beta_1 = 0$)
- H_α: There is some relationship ($\beta_1 \neq 0$)
- t-statistic

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

- t values of larger magnitudes (either negative or positive) are less likely. This probability is known as p-value. A small p-value indicates that it is unlikely to observe such a substantial association between the predictor and the response due to chance

P-VALUE OF THE STATISTICS



ADVERTISING DATA

	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

We reject the null hypothesis
Alternative hypothesis is true
TV and Sales are related

ACCURACY OF THE MODEL - R^2

- R^2 statistic

$$R^2 = \frac{variance\ explained}{variance\ total} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}, \qquad 0 \leq R^2 \leq 1$$

$$TSS = \sum_{k=1}^{n} (y_k - \bar{y})^2, \qquad RSS = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2$$

- In the simple linear regression setting

$$R^2 = r^2$$

where r is the correlation coefficient between X and Y

ACCURACY OF THE MODEL - RSE

- Residual standard error (RSE)

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2}\sum_{k=1}^{n}(y_k - \hat{y}_k)^2}$$

- RSE is measured in the units of Y

ADVERTISING DATA

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

- Actual sales in each market deviate from the true regression line by approximately $3{,}260$ units, on average
- In the advertising data set, the mean value of sales over all markets is approximately $14{,}000$ units, and so the percentage error is

$$3{,}260/14{,}000 = 23\%$$

- $R^2 = 0.61$ in the $sales{\sim}TV$ model, so $two-thirds$ of the variability in sales is explained by a linear regression on TV

ADVERTISING DATA

Simple regression of `sales` on `radio`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	9.312	0.563	16.54	< 0.0001
`radio`	0.203	0.020	9.92	< 0.0001

Simple regression of `sales` on `newspaper`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	< 0.0001

MULTIPLE LINEAR REGRESSION

MULTIPLE LINEAR REGRESSION

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + e$$

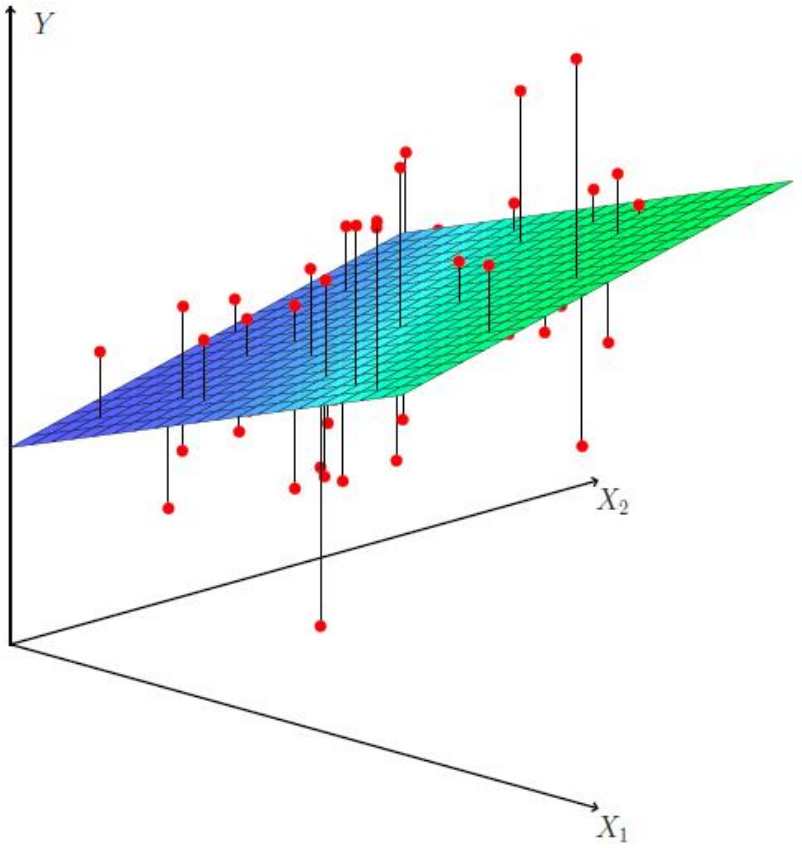
$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

$$sales = \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 newspaper + e$$

$$\widehat{sales} = \hat{\beta}_0 + \hat{\beta}_1 TV + \hat{\beta}_2 radio + \hat{\beta}_3 newspaper$$

$$RSS = \sum_{k=1}^{n} (y_k - \hat{y}_k)^2 \rightarrow min$$

MULTIPLE LINEAR REGRESSION



COEFFICIENTS ESTIMATE

$$y_j = \sum_{k=0}^{p} \beta_k x_{k,j} + e$$

$$I(\beta) = \sum_{j=1}^{n} \left(y_j - \sum_{k=0}^{p} \beta_k x_{k,j} \right)^2 \to min\,, \qquad \beta = (\beta_0, \dots, \beta_p)^T$$

$$\frac{\partial I}{\partial \beta_m} = -2 \sum_{j=1}^{n} \left[y_j - \sum_{k=0}^{p} \beta_k x_{k,j} \right] x_{m,j} = 0, \qquad m = 0, \dots, p$$

COEFFICIENTS ESTIMATE

$$\sum_{j=1}^{n}\left[y_j - \sum_{k=0}^{p}\beta_k x_{k,j}\right]x_{m,j} = 0, \qquad m = 0, \dots, p$$

$$\sum_{j=1}^{n} y_j x_{m,j} = \sum_{j=1}^{n} x_{m,j} \sum_{k=0}^{p}\beta_k x_{k,j}\,, m = 0, \dots, p$$

$$b = (y_1, y_2, \dots, y_N)^T, \qquad M = (x_{m,j})$$

$$M^T M\beta = M^T b$$

- Normal system, always solvable - the solution is not always unique

HESSIAN MATRIX FOR A LINEAR REGRESSION

$$\frac{\partial^2 I}{\partial \beta_m \partial \beta_r} = H_{mr} = 2\sum_{j=1}^{n} x_{m,j} x_{r,j}\,, \qquad r, m = 0, \ldots, p$$

$$H = MM^T$$

which is always positive definite if matrix M has full rank (all columns are independent)

SALES ~ TV, RADIO, NEWSPAPER

close to simple regression coefficients

	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

How we interpret the results? For a given amount of TV and newspaper advertising, spending an additional $1,000 on radio advertising leads to an increase in sales by approximately 189 units.

SALES ~ TV, RADIO, NEWSPAPER

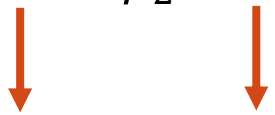
	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

- The simple and multiple regression coefficients can be quite different
- The problem is in correlations between different features
- Simple linear regression is unable to detect that correlation

THE DIFFERENCE BETWEEN REGRESSIONS

- Simple regression

$$sales = \beta_0 + \beta_1 TV + \beta_2\ Radio + \beta_3\ Newspaper$$



0 0

- Multiple regression

$$sales = \beta_0 + \beta_1 TV + \beta_2\ Radio + \beta_3\ Newspaper$$

is fixed is fixed

CRAZY EXAMPLE

- Regression of shark attacks versus ice cream sales at a given beach would show a positive relationship (probably)
- However, we can not reduce the shark attacks by banning the ice creams at beaches
- Ice cream sales doesn't impact the shark attacks

IS THERE A RELATIONSHIP BETWEEN THE RESPONSE AND PREDICTORS?

- Is at least one of the predictors useful?
- Null hypothesis

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

- Alternative hypothesis

$$H_\alpha : \ at\ least\ one\ \beta_j\ is\ non - zero$$

- F-statistic

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

- If H_0 is true $F \approx 1$, otherwise $F \gg 1$
- In our model $\boldsymbol{F = 570}$ and p-value is almost **0**. Hence, we have extremely strong evidence that at least one of the media is associated with increased sales

DO ALL PREDICTORS HELP TO EXPLAIN Y?

- Although there are approaches to answer the question
  - Forward selection
  - Backward selection
  - Mixed selection
- The best answer is possible to get by lasso and ridge regression

MEASURES FOR THE QUALITY OF MODELS

- Adjusted R^2
- AIC (Akaike information criterion)
- BIC (Bayesian information criterion)

These methods add penalty to RSS for the number of variables (i.e. complexity)

MEASURES FOR THE QUALITY OF MODELS

For a fitted least squares model containing p predictors and with σ^2 estimated variance of the error e

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

$$AIC = \frac{1}{n\sigma^2}(RSS + 2p\sigma^2)$$

$$BIC = \frac{1}{n}(RSS + \log(n)\, p\sigma^2)$$

MODEL ACCURACY

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

- As closer is R^2 to 1 as better

$$RSE = \sqrt{\frac{1}{n-p-1} RSS}$$

- As smaller is the RSE as better. It is measured in the units of Y and it is not always clear what is a good RSE

SALES ~ TV, RADIO, NEWSPAPER

- The model containing only TV as a predictor has

$$R^2 = 0.61, \qquad RSE = 3.26$$

- The model containing TV and radio has an

$$R^2 = 0.89719, \qquad RSE = 1.681$$

- The model with TV, radio and newspaper has

$$R^2 = 0.8972, \qquad RSE = 1.686$$

- Newspaper variable can be dropped from the model

ACCURACY MEASURES FOR ADVERTISING DATA

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

UNCERTAINTIES IN THE MODEL

- Three sorts of uncertainty:
  - Coefficients estimates – confidence intervals
  - Linear model might not be correct
  - Irreducible error – prediction intervals

CONFIDENCE INTERVAL

- 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 $f(X)$

PREDICTION INTERVAL

- The response value cannot be predicted perfectly because of the random error in the model

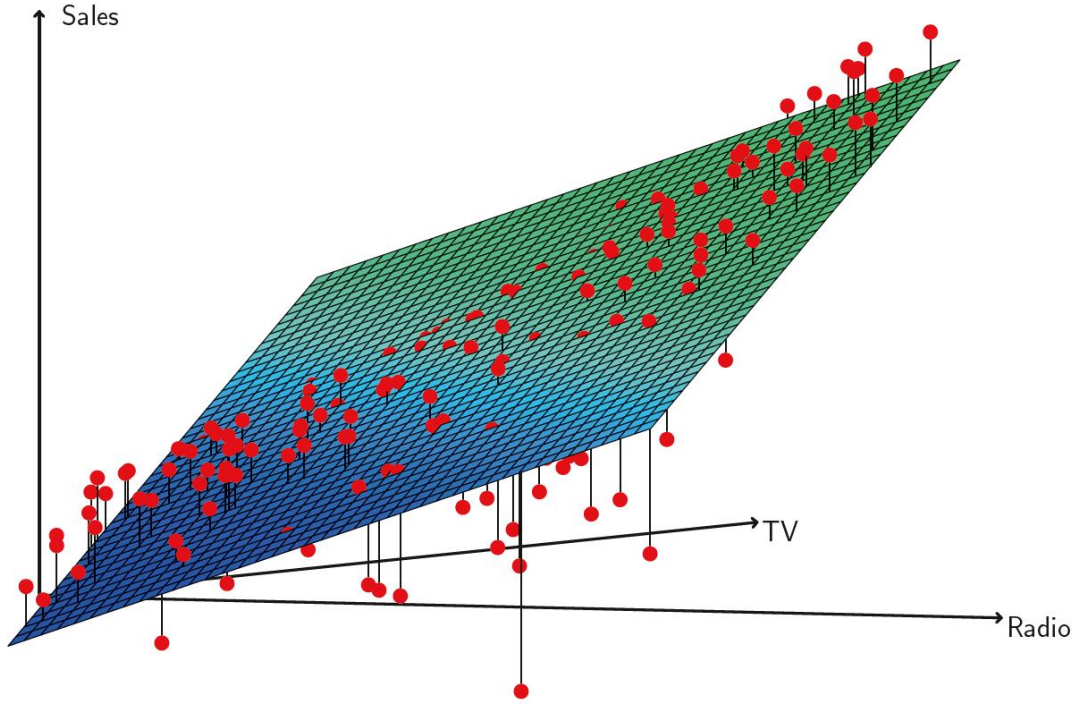
$$Y = f(X) + e$$

- We use ***prediction intervals*** to determine how close $\hat{Y}$ will be to Y
- Prediction intervals are always wider than confidence intervals, because they incorporate both the reducible and irreducible errors

ADVERTISING DATA

- Assume spending 100,000$ in TV and 20,000$ in radio
- Predicted exact value is 11.25
- Confidence interval is [10.985; 11.528]
- Prediction interval is [7.930; 14.580]

MARKETING SUMMARY - SYNERGY



- We observe non-linear effects and there is a synergy among the advertising media

INCLUDING NON-LINEAR (INTERACTION) TERMS

$$sales = \beta_0 + \beta_1\ TV + \beta_2\ radio + \beta_3\ (radio \times TV) + e$$

	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

INCLUDING NON-LINEAR (INTERACTION) TERMS

- These results strongly suggest that the model that includes the interaction term is superior to the model that contains only *main effects*
- The p-value for the interaction term, TV×radio, is extremely low
- It turned out 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
- In this situation, given a fixed budget of $100,000, spending half on radio and half on TV may increase sales more than allocating the entire amount to either TV or to radio (*synergy* effect)

MARKETING SUMMARY

1. Is there a relationship between advertising sales and budget?
   - We have clear evidence of a relationship between advertising and sales (based on F-statistic as the corresponding p-value is very small).

2. How strong is the relationship?
   - Relationship is rather strong. $RSE = 1681$ units while the mean value of the response is 14022 units, indicating a percentage error of 12%.
   - $R^2 = 0.9$ indicating that almost 90% of the variance can be explained by the linear model.

MARKETING SUMMARY

3. Which media contribute to sales?
   - Only TV and radio are related to sales (according to the corresponding p-values).
4. How large is the effect of each medium on sales?
   - Standard errors of each coefficient can be used
   - TV - $(0.043,0.049)$, Radio - $(0.172,0.206)$, Newspaper - $(-0.013,0.011)$
   - The confidence intervals for TV and Radio are far from 0 and narrower showing that this media are related to sales.
   - Simple regression should help to estimate the impact of each medium
   - Strong association between TV and sales and between the Radio and sales

MARKETING SUMMARY

5. How accurately can we predict future sales?
   - Individual response $f(x) + e$ or the average response $f(x)$?
   - We use prediction interval or confidence interval, respectively

6. Is the relationship linear?
   - No
   - There are non-linear effects in the residuals (based on the plot)

7. Is there synergy among the media?
   - Yes
   - Including an interaction term in the model results in substantial increase in R^2, from around 90% to almost 97%

QUALITATIVE PREDICTORS

- We create an indicator or *dummy variable* that takes on two possible dummy numerical values
- Example: male, female

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

$$y_i = \beta_0 + \beta_1 x_i + e_i = \begin{cases} \beta_0 + \beta_1 + e_i, & if\ i^{th}\ person\ is\ female \\ \beta_0 + e_i, & if\ i^{th}\ person\ is\ male \end{cases}$$

QUALITATIVE PREDICTORS

- Example: Asian, Caucasian, African American

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

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

POTENTIAL PROBLEMS DIAGNOSTIC PLOTS

POTENTIAL PROBLEMS

- Non-linearity of the response-predictor relationship
- Non-constant variance of residuals
- Normality of residuals
- Correlation of residuals
- Influential points
  - Outliers
  - High-leverage points
- Collinearity

DIAGNOSTIC PLOTS

- "Residuals vs Fitted Values"
- "Normal Q-Q" plot
- "Scale-Location" or "Spread-Location" graph
- "Residuals vs Leverage"

NON-LINEARITY

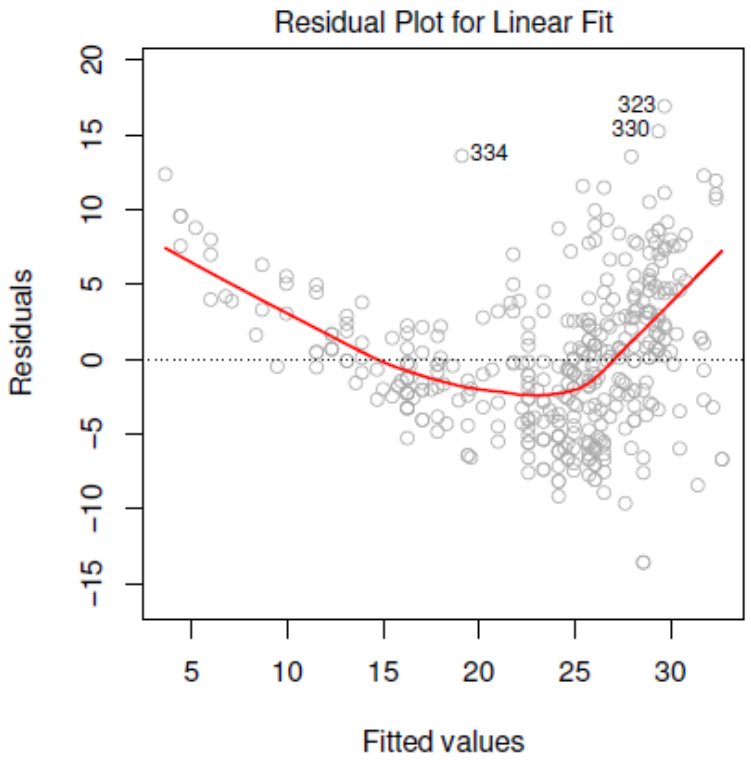
NON-LINEARITY

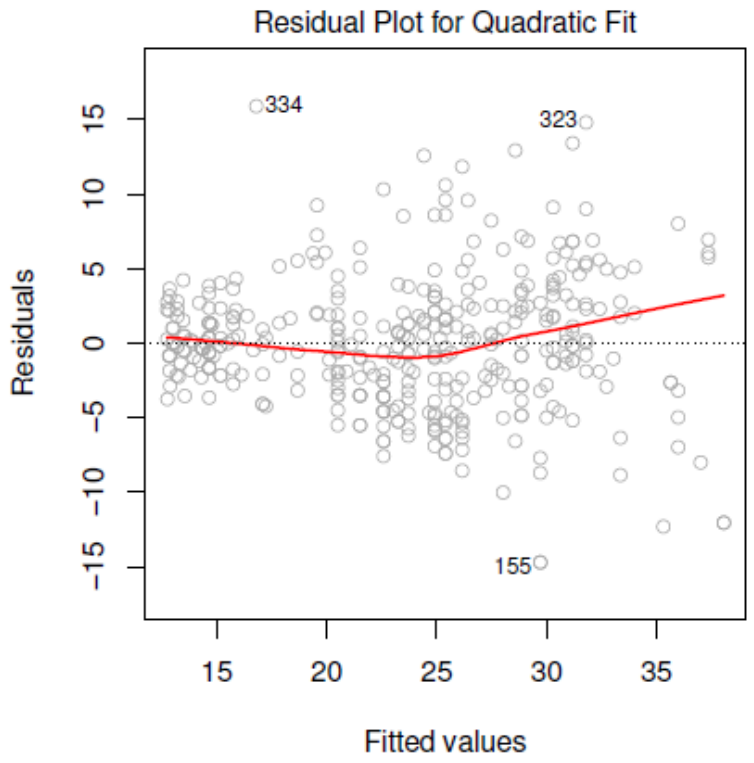
- The linear regression model assumes that there is a straight-line relationship between the predictors and the response
- If the true relationship is not linear, then all our conclusions are suspicious
- Solution: Look into the "**residuals vs fitted values**" graphs for identifying non-linearity

RESIDUALS VS FITTED VALUES

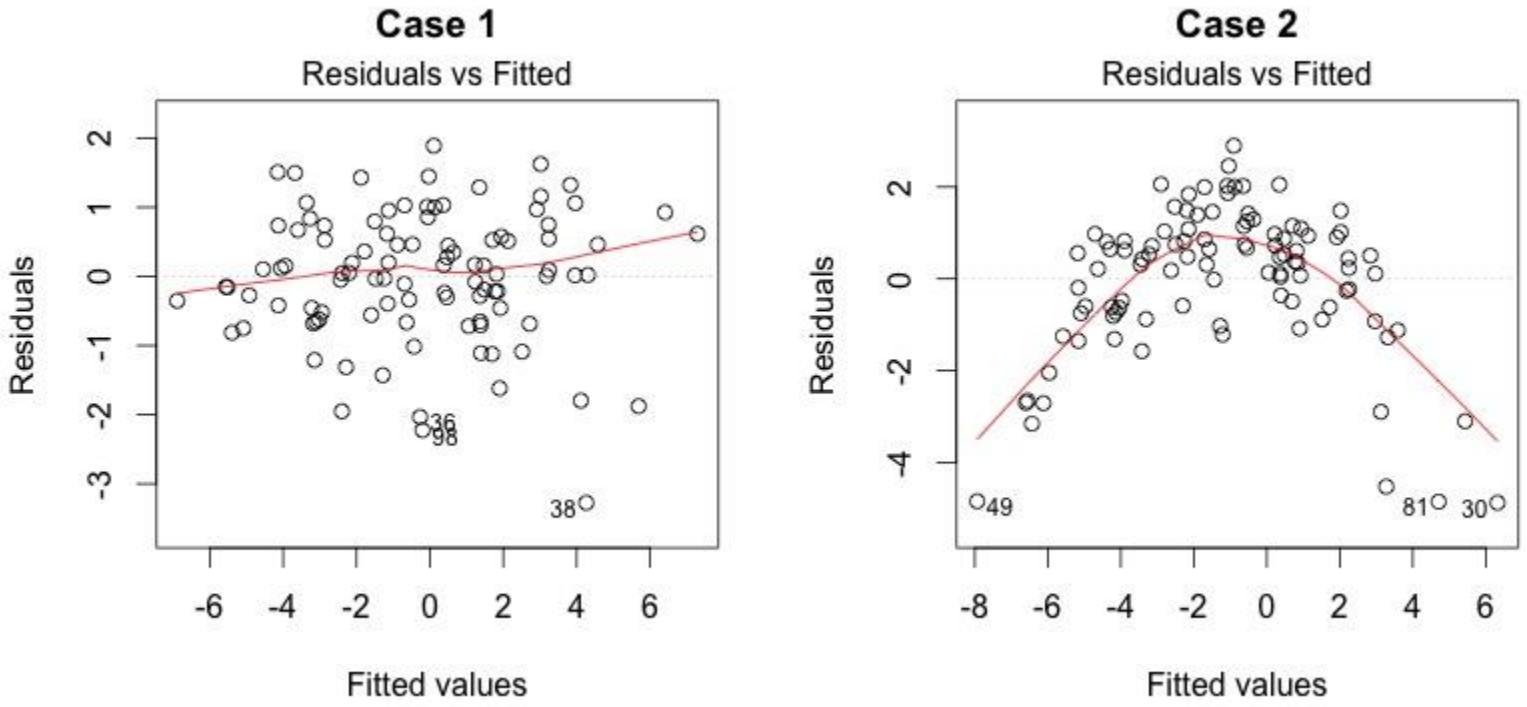
- This plot shows if residuals have non-linear patterns
- There could be a non-linear relationship between predictor variables and an outcome variable and the pattern could show up in this plot if the model doesn't capture the non-linear relationship
- If you find equally spread residuals around a horizontal line without distinct patterns, that is a good indication you don't have non-linear relationships

NON-LINEARITY





NON-LINEARITY



- There is not any distinctive pattern in Case 1
- We see a parabola in Case 2
- In Case 2, the non-linear relationship was not explained by the model and was left out in the residuals

NON-CONSTANT VARIANCE

NON-CONSTANT VARIANCE

- Another important assumption of the linear regression model is that the error terms have a constant variance,

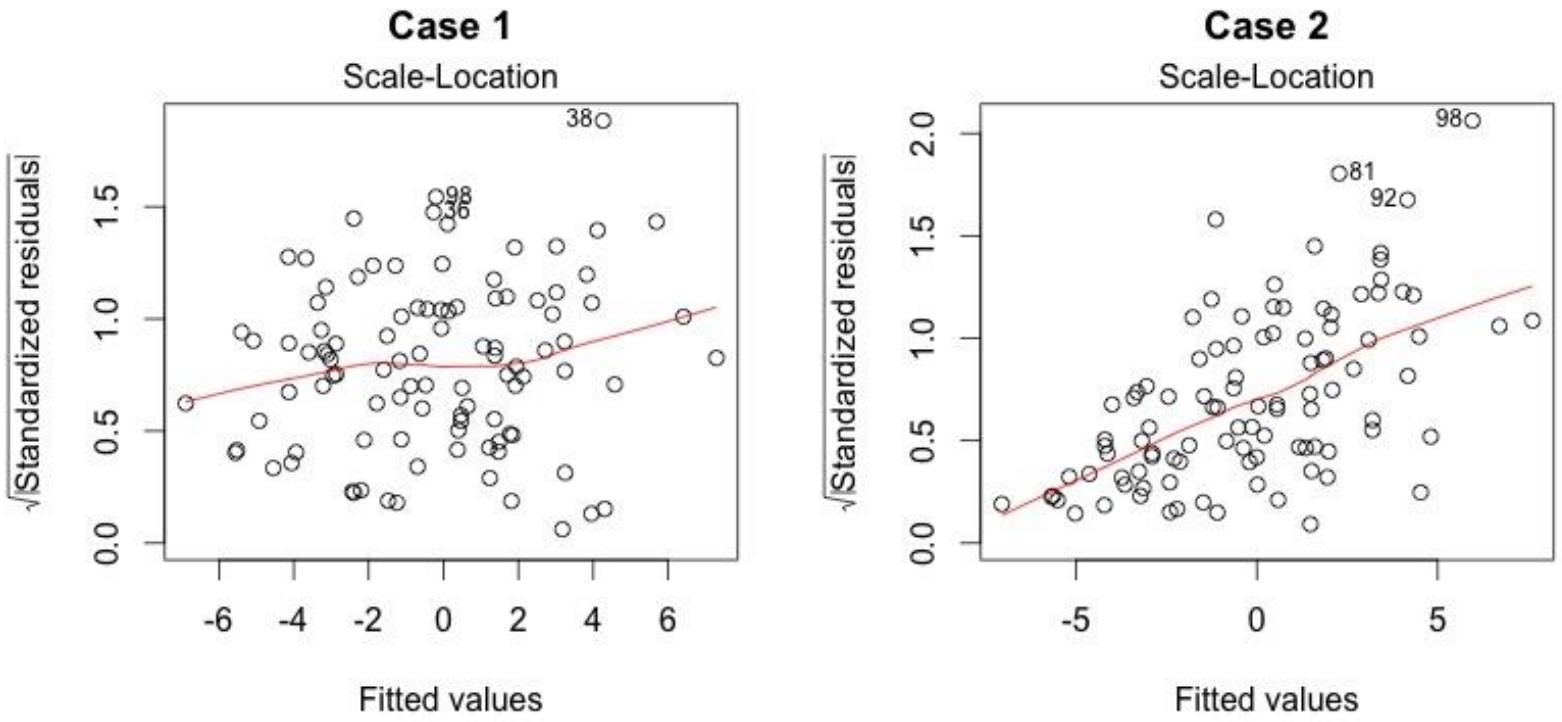
$$Var(e_i) = \sigma^2$$

- The standard errors, confidence intervals, and hypothesis tests associated with the linear model rely upon this assumption
- Solution: Look into the "scale-location" ("spread-location") graph for identifying non-constant variance

SCALE-LOCATION PLOT

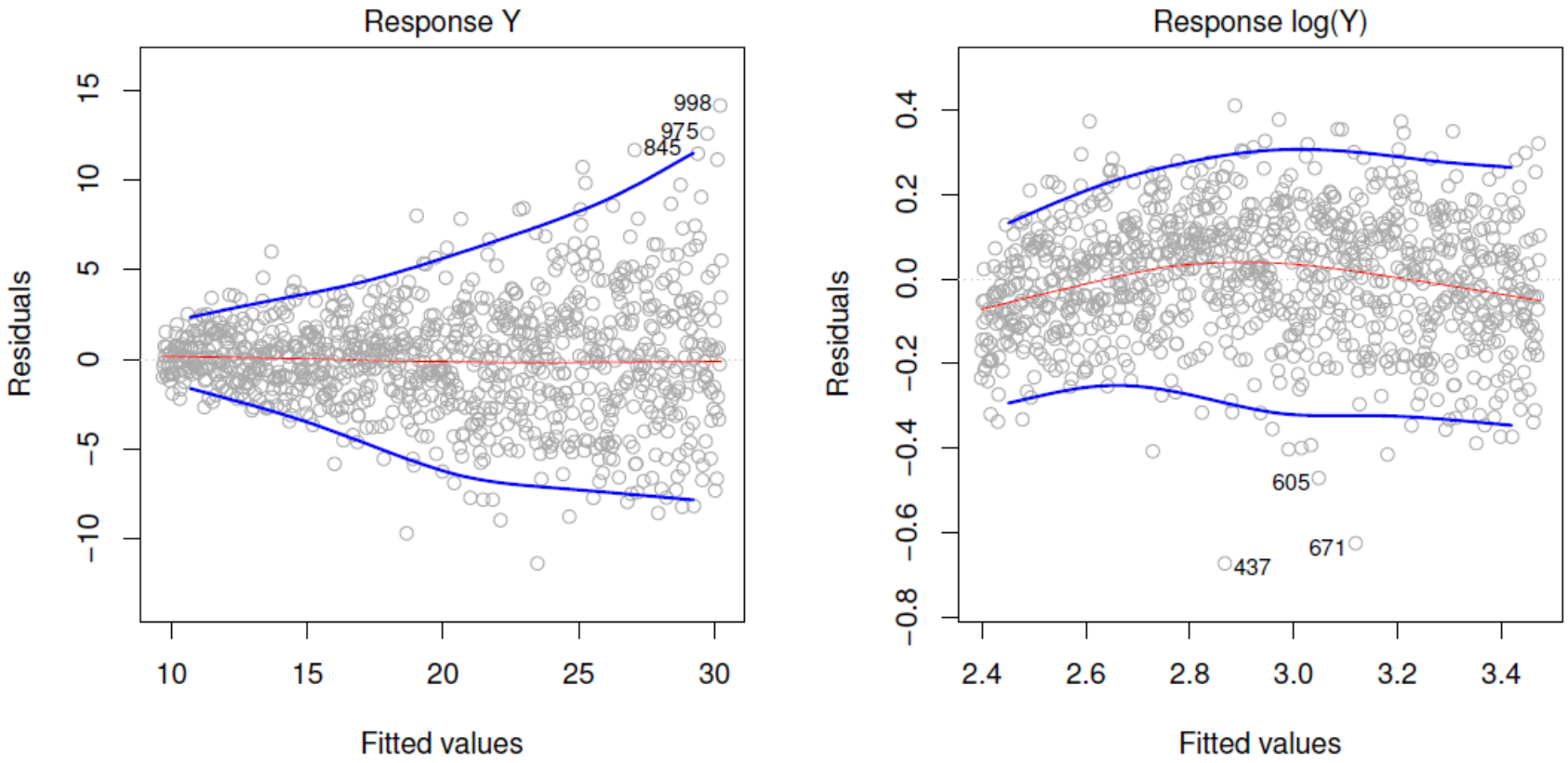
- This plot shows if residuals are spread equally along the ranges of predictors
- This is how you can check the assumption of equal variance (homoscedasticity)
- It's good if you see a horizontal line with equally (randomly) spread points

SCALE-LOCATION



- In Case 1, the residuals appear randomly spread
- In Case 2, the residuals begin to spread wider along the x-axis as it passes around 5
- Because the residuals spread wider and wider, the red smooth line is not horizontal and shows a steep angle in Case 2

NON-CONSTANT VARIANCE OF ERROR

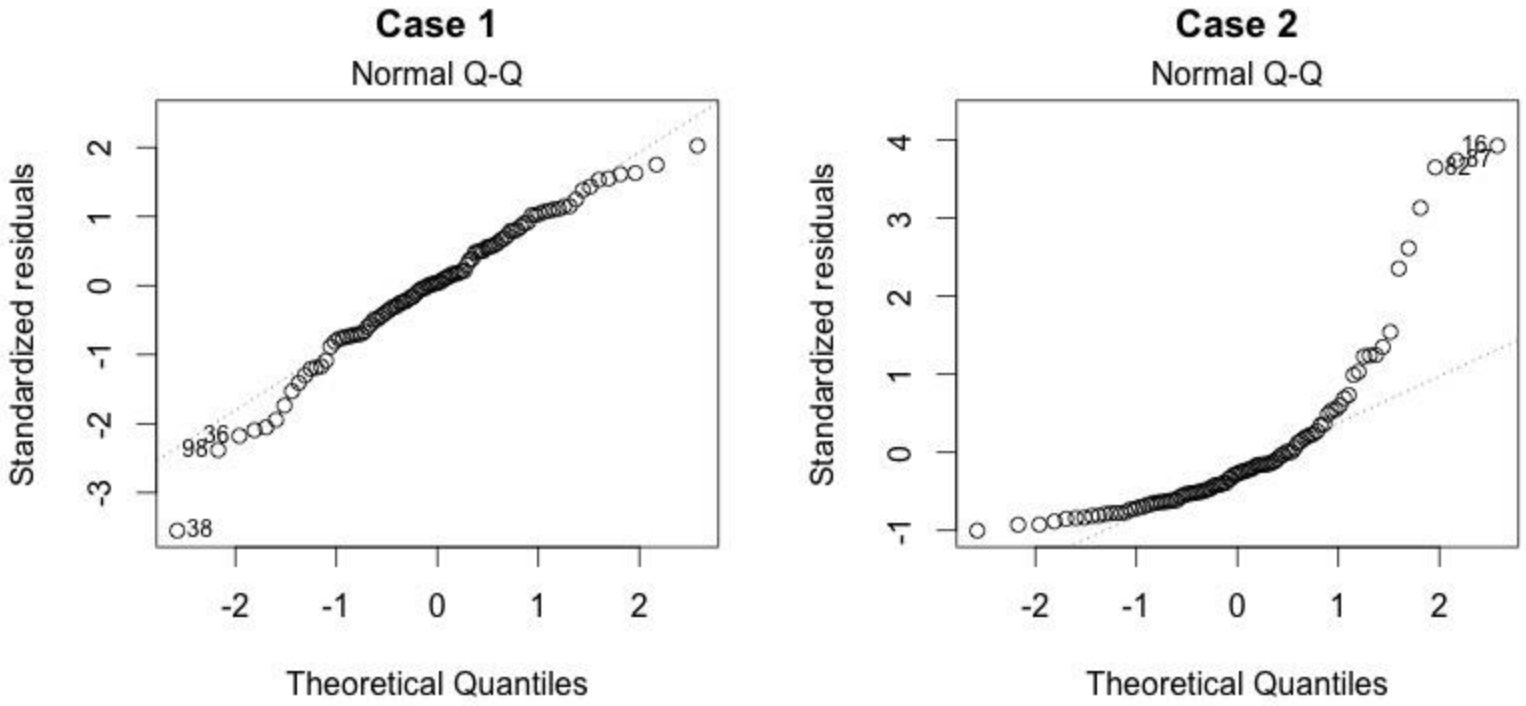


NORMALITY

NORMALITY OF RESIDUALS

- Residuals should follow an i.i.d. process. They need to be independent (uncorrelated) and have the same normal distribution
- Normality can be checked by the well-known Q-Q plot
- This plot shows if residuals are normally distributed
- Do residuals follow a straight line well or do they deviate severely?
- It's good if residuals are lined well on the straight dashed line

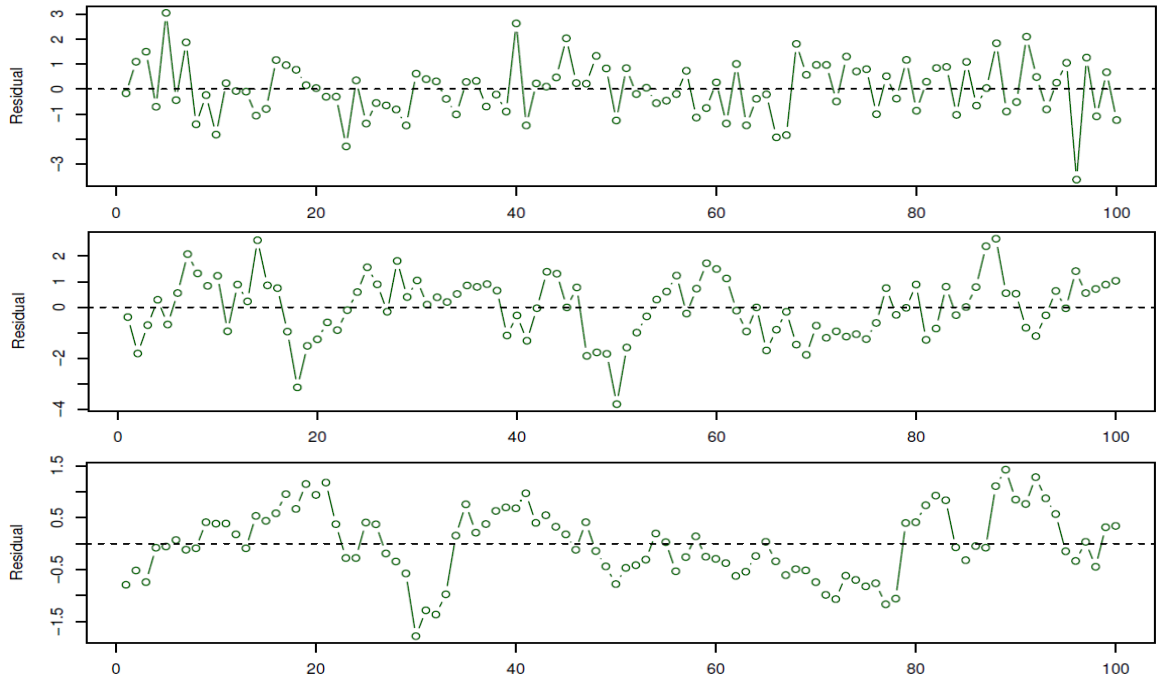
Q-Q PLOT



- Of course they wouldn't be a perfect straight line
- Case 2 definitely concerns
- Case 1 doesn't concern too much, although an observation numbered as 38 looks a little off.
- Let's look at the next plots while keeping in mind that #38 might be a potential problem.

CORRELATION OF ERROR TERMS

- Linear regression assumes that the error terms e_i are uncorrelated
- For instance, the fact that e_i is positive provides no information about the sign of e_{i+1}

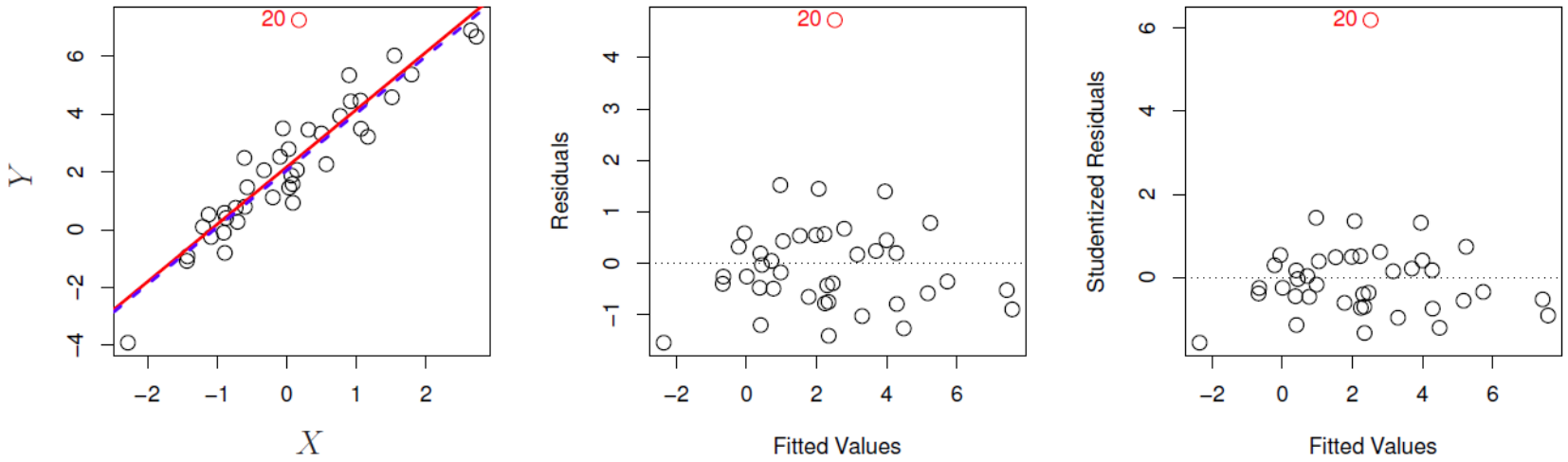


INFLUENTIAL POINTS

OUTLIERS

- ***Outlier*** is a point for which y_i is far from the value predicted by the model
- Observations whose standardized residuals are greater than 3 in absolute value are possible outliers

OUTLIERS



- **Left:** The least squares regression line is shown in red. The regression line after removing the outlier is shown as dashed line
- **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

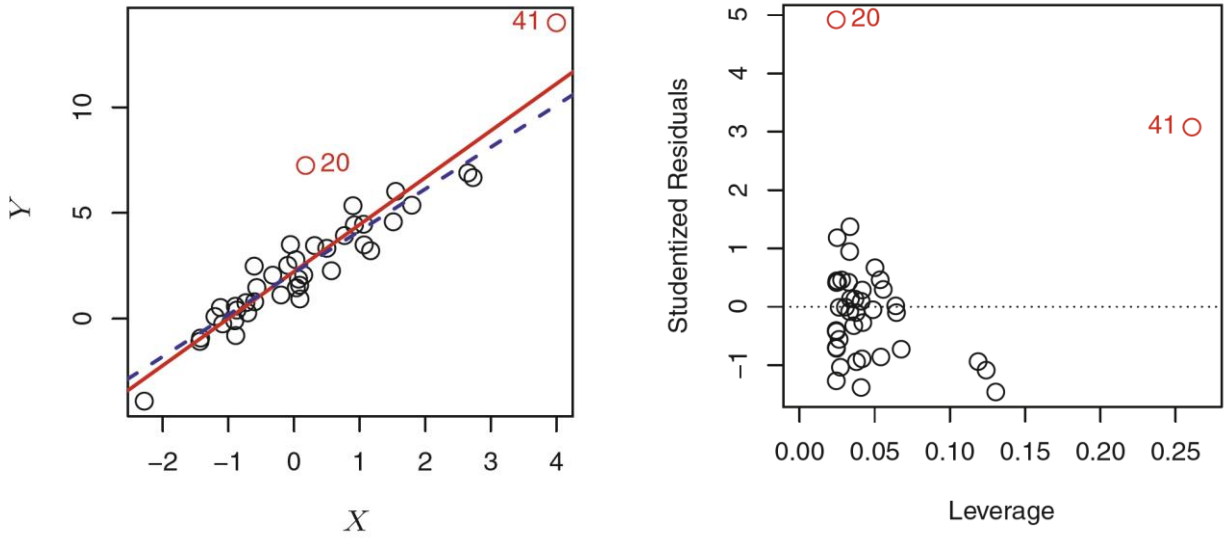
HIGH-LEVERAGE POINTS

- Outliers were the observations with unusual response
- In contrast, observations with high leverage have an unusual input
- Solution: compute the *leverage statistic* and remove the points with the large values
- Leverage statistic for simple regression

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

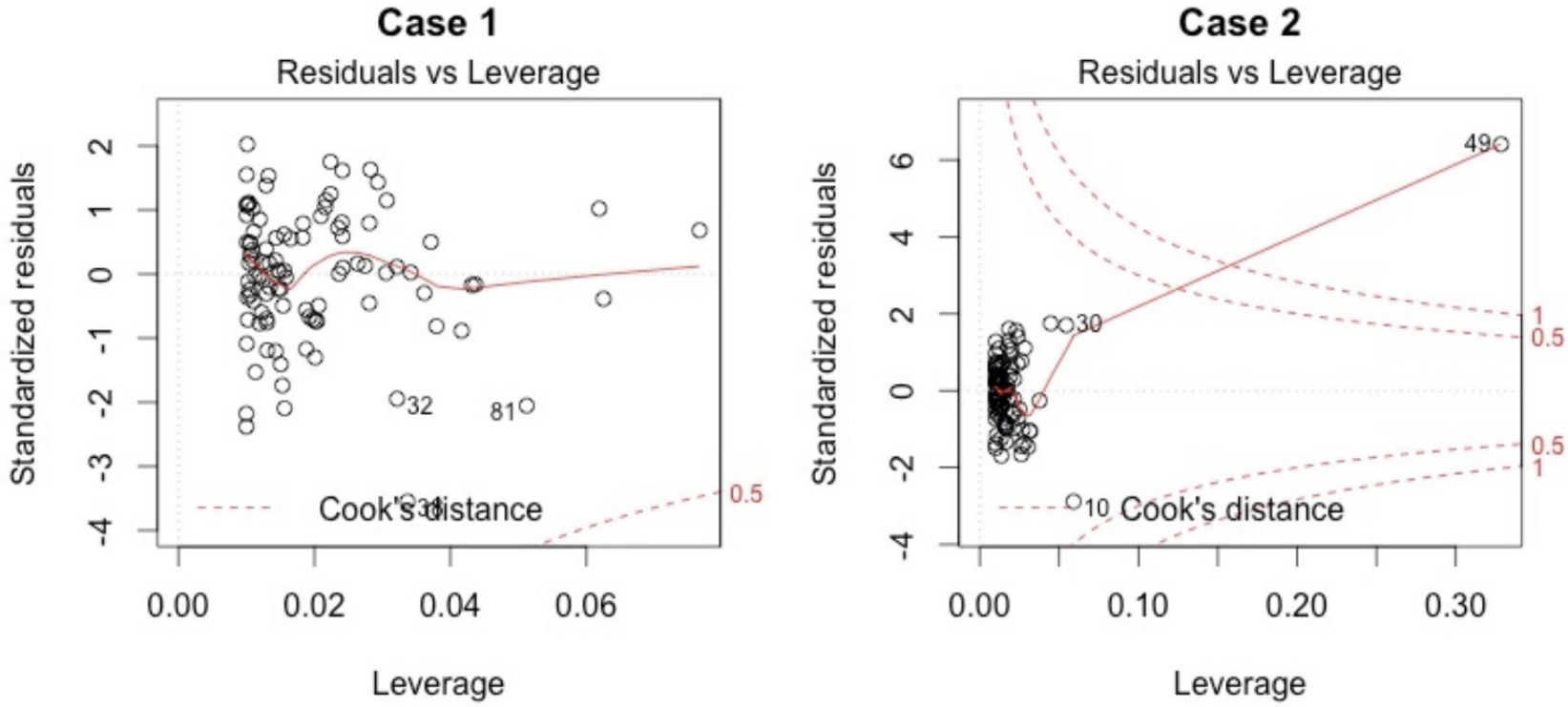
- If h_i is big then, the point has high leverage

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 blue line is the fit with observation 41 removed
- Right: Observation 41 has a high leverage and a high residual. It is an outlier as well as a high leverage observation. This is a particularly dangerous combination!

INFLUENTIAL POINTS



- Case 1 is the typical look when there is no influential case, or cases. You can barely see Cook's distance lines (a red dashed line) because all cases are well inside of the Cook's distance lines.
- In Case 2, a case is far beyond the Cook's distance lines (the other residuals appear clustered on the left because the second plot is scaled to show larger area than the first plot). The plot identified the influential observation as #49.

COLLINEARITY

COLLINEARITY

- *Collinearity* means that two or more predictor variables are correlated
- The presence of collinearity poses problems, since it can be difficult to separate out the individual effects of collinear variables on the response
- Solution: Calculate *variance inflation factor (VIF)* and remove predictors with high VIF

COLLINEARITY

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

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
- As a rule, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity
- The smallest value is 1 indicating complete absence of collinearity

INTERESTING EXAMPLE

ANSCOMBE'S QUARTET