Linear Regression: Extensions

Damek Davis
School of ORIE, Cornell University
ORIE 4740 Lec 4 (Jan 30)

Announcements

Recap: Model Evaluation

- Accuracy of Model:
  - Residual Sum of Squares: RSS $\triangleq \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.
  - Residual Standard Error: RSE $\triangleq \sqrt{\frac{1}{n-p-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$.
  - R^2 statistics: $R^2 = \frac{\text{TSS}-\text{RSS}}{\text{TSS}} = \frac{\text{explained var.}}{\text{total var.}}$
  - Hypothesis test of the whole model:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \quad \text{vs.} \quad H_1 : \text{at least one } \beta_j \text{ is non-zero}$$

    Smaller p-value $\Rightarrow$ More evidence of relationship (against H_0).
    (Standard cutoff: 0.05, 0.01)

Recap: Model Evaluation

- Accuracy of Model:
  - Residual Sum of Squares: RSS $\triangleq \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$.
  - Residual Standard Error: RSE $\triangleq \sqrt{\frac{1}{n-p-1} \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}$.
  - R^2 statistics: $R^2 = \frac{\text{TSS}-\text{RSS}}{\text{TSS}} = \frac{\text{explained var.}}{\text{total var.}}$
  - Hypothesis test of the whole model:

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \quad \text{vs.} \quad H_1 : \text{at least one } \beta_j \text{ is non-zero}$$

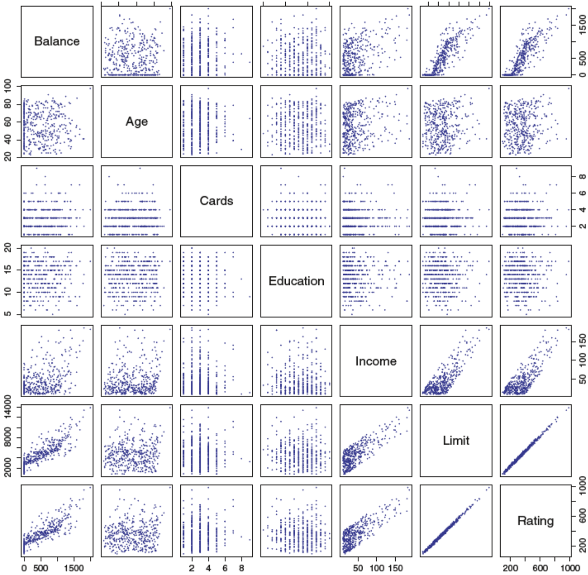
  Smaller p-value $\Rightarrow$ More evidence of relationship (against H_0).
  (Standard cutoff: 0.05, 0.01)

- Accuracy of Coefficient Estimates:
  - Standard Error (SE) of each estimate: SE$(\hat{\beta}_j)$
  - An estimate of $|\hat{\beta}_j - \beta_j|$
  - Can be used to construct confidence intervals.
  - Can be used to perform hypothesis tests

Today: Other considerations in linear regression

Reading: ISLR Section 3.3

Categorical Predictors



Categorical Predictors

(ISLR Section 3.3.1)

- Married: Yes, No
- Fuel Type: Diesel, Petrol, CNG

Predictors with 2 levels:

Often called dummy variables.

Predictors with $>$ 2 levels

Categorical Predictors

(ISLR Section 3.3.1)

- Married: Yes, No
- Fuel Type: Diesel, Petrol, CNG

Predictors with 2 levels:

$$x_{i1} = \begin{cases} 1 & \text{if person is married} \\ 0 & \text{if person is not married} \end{cases}$$

Often called dummy variables.

Predictors with $>$ 2 levels

Categorical Predictors

(ISLR Section 3.3.1)

- Married: Yes, No
- Fuel Type: Diesel, Petrol, CNG

Predictors with 2 levels:

$$x_{i1} = \begin{cases} 1 & \text{if person is married} \\ 0 & \text{if person is not married} \end{cases}$$

Often called dummy variables.

Predictors with $>$ 2 levels

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person uses Diesel} \\ 0 & \text{otherwise} \end{cases}, \quad \text{and} \quad x_{i3} = \begin{cases} 1 & \text{if } i\text{th person uses Petrol} \\ 0 & \text{otherwise} \end{cases}$$

If $x_{i2} = x_{i3} = 0$, then fuel type is CNG.

Categorical Predictors

Predictors with 2 levels:

Predictors with $>$ 2 levels

Categorical Predictors

Predictors with 2 levels:

$$\text{balance}(i) = \beta_0 + \beta_1 \text{Married}(i) + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is married} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is not married} \end{cases}$$

Predictors with $>$ 2 levels

Categorical Predictors

Predictors with 2 levels:

$$\text{balance}(i) = \beta_0 + \beta_1 \text{Married}(i) + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is married} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is not married} \end{cases}$$

Predictors with >2 levels

$$\begin{aligned} \text{balance}(i) &= \beta_0 + \beta_1 \text{Diesel}(i) + \beta_2 \text{Petrol}(i) + \epsilon_i \\ &= \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person uses Diesel} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person uses Petrol} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person uses CNG} \end{cases} \end{aligned}$$

Nonlinearity

🔥 **Kareem Carr** 🔥
@kareem_carr

God of Statistics: *creates linear regression* people can use you when stuff is linear

linear regression: what about when stuff isn't linear?

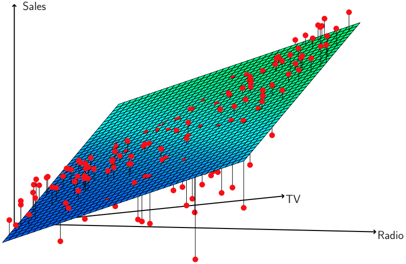
God of Statistics: shhh. we don't talk about that

8:59 PM · Jul 17, 2020 · Twitter for iPhone

46 Retweets **5** Quote Tweets **488** Likes

Nonlinearity: Advertising Example

TV	Radio	Newspaper	Sales
230.1	37.8	69.2	22.1
44.5	39.3	45.1	10.4
17.2	45.9	69.3	9.3
151.5	41.3	58.5	18.5
180.8	10.8	58.4	12.9
8.7	48.9	75.0	7.2
57.5	32.8	23.5	11.8
120.2	19.6	11.6	13.2
8.6	2.1	1.0	4.8
199.8	2.6	21.2	10.6
66.1	5.8	24.2	8.6
214.7	24.0	4.0	17.4
23.8	35.1	65.9	9.2
97.5	7.6	7.2	9.7
204.1	32.9	46.0	19.0



Assume a linear model:

$$\text{sales}(i) \approx \beta_0 + \beta_1 \times \text{TV}(i) + \beta_2 \times \text{radio}(i) + \beta_3 \times \text{newspaper}(i),$$

$$i = 1, 2, \ldots, n$$

Nonlinearity

(ISLR Section 3.3.2 and 3.3.3)

Original Model:

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \epsilon$$

Nonlinearity

(ISLR Section 3.3.2 and 3.3.3)

Original Model:

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \epsilon$$

Interaction terms:

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

Nonlinearity

(ISLR Section 3.3.2 and 3.3.3)

Original Model:

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \epsilon$$

Interaction terms:

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

Transform the Data Table:

1	TV	radio	TV$\times$ radio
1	.83	.3	0.249
$\vdots$	$\vdots$	$\vdots$	$\vdots$

Nonlinearity

(ISLR Section 3.3.2 and 3.3.3)

Original Model:

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \epsilon$$

Interaction terms:

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

Transform the Data Table:

1	TV	radio	TV$\times$ radio
1	.83	.3	0.249
$\vdots$	$\vdots$	$\vdots$	$\vdots$

	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`$\times$`radio`	0.0011	0.000	20.73	< 0.0001

- R^2 : 89.7% $\rightarrow$ 96.8%
- Applies to categorical predictors as well.

Nonlinearity

(ISLR Section 3.3.2 and 3.3.3)

Original Model:

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \epsilon$$

Interaction terms:

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

Transform the Data Table:

1	TV	radio	TV$\times$ radio
1	.83	.3	0.249
$\vdots$	$\vdots$	$\vdots$	$\vdots$

	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$\times$radio	0.0011	0.000	20.73	< 0.0001

Interpretation: A $1000 increase in TV results in increased sales of

$$19 + 1.1 \times \text{radio units}$$

Nonlinearity

(ISLR Section 3.3.2 and 3.3.3)

Original Model:

$$\text{sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \epsilon$$

Interaction terms:

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

Transform the Data Table:

1	TV	radio	TV$\times$ radio
1	.83	.3	0.249
$\vdots$	$\vdots$	$\vdots$	$\vdots$

	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$\times$radio	0.0011	0.000	20.73	< 0.0001

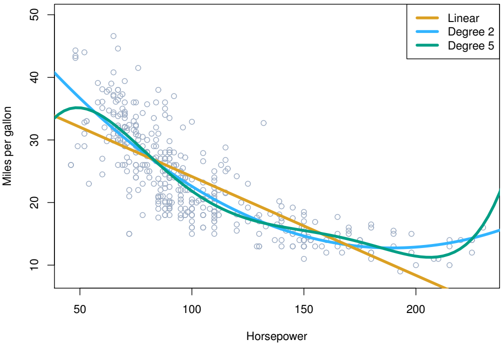
A \$1000 increase in radio results in increased sales of

A. $19 + 1.1 \times$ TV units.

B. $289 + 1.1 \times$ TV units

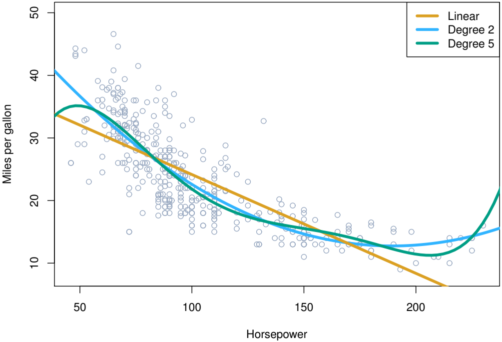
Nonlinearity

High order terms:



Nonlinearity

High order terms:

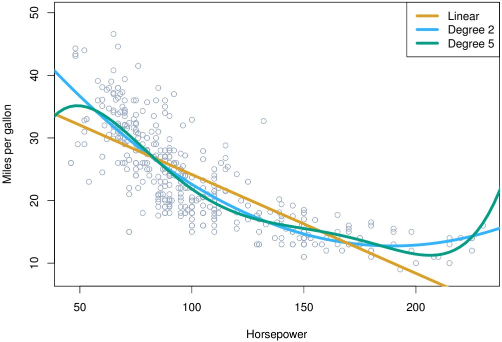


$$\text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times \text{horsepower}^2 + \epsilon$$

- R^2 : 60.6% $\rightarrow$ 68.8%

Nonlinearity

High order terms:

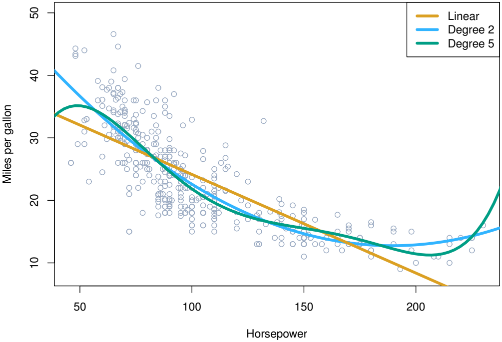


$$\text{mpg} = \beta_0 + \beta_1 \times \text{horsepower} + \beta_2 \times \text{horsepower}^2 + \epsilon$$

- R^2 : 60.6% $\rightarrow$ 68.8%
- Can include higher order terms. Danger of overfitting!

Nonlinearity

High order terms:



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

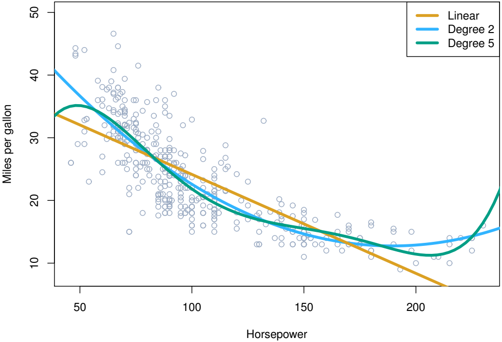
How do we find the the coefficients of this model?

A. with linear algebra

B. with new techniques we haven't covered yet

Nonlinearity

High order terms:



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

Transform the Data Table:

1	horsepower	horsepower^2
1	120	14,400
$\vdots$	$\vdots$	$\vdots$

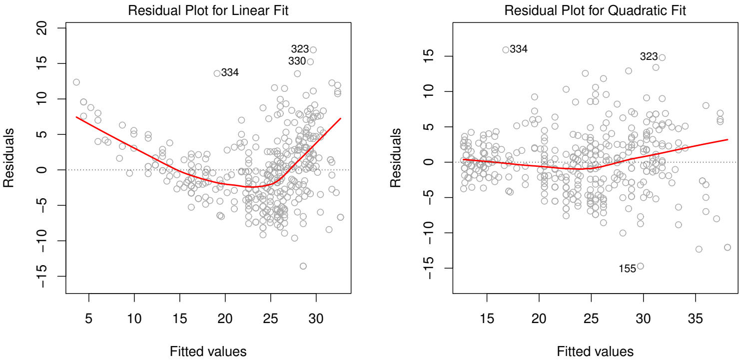
Detecting Nonlinearity

Detecting Nonlinearity

Residual plots: $y_i - \hat{y}_i$ versus $\hat{y}_i$

Detecting Nonlinearity

Residual plots: $y_i - \hat{y}_i$ versus $\hat{y}_i$



Nonlinearity: Summary

Interaction terms: $X_1X_2,\ X_2X_5X_6, \ldots$

High order terms: $X_1^2,\ X_1^2,\ X_1^3, \ldots$

Nonlinearity: Summary

Interaction terms: $X_1 X_2, X_2 X_5 X_6, \ldots$

High order terms: $X_1^2, X_1^2, X_1^3, \ldots$

Other functional terms: $\sqrt{X_2}, \log X_2, \ldots$

Nonlinearity: Summary

Interaction terms: X_1X_2, $X_2X_5X_6$, $\ldots$

High order terms: X_1^2, X_1^2, X_1^3, $\ldots$

Other functional terms: $\sqrt{X_2}$, $\log X_2$, $\ldots$

Detecting nonlinearity: Residual plots

Nonlinearity: Summary

Interaction terms: $X_1 X_2$, $X_2 X_5 X_6$, ...

High order terms: X_1^2, X_1^2, X_1^3, ...

Other functional terms: $\sqrt{X_2}$, $\log X_2$, ...

Detecting nonlinearity: Residual plots

Caution: Overfitting!

Nonlinearity: Summary

Interaction terms: $X_1X_2, X_2X_5X_6, \dots$

High order terms: $X_1^2, X_1^2, X_1^3, \dots$

Other functional terms: $\sqrt{X_2}, \log X_2, \dots$

Detecting nonlinearity: Residual plots

Caution: Overfitting!

More on nonlinear methods later (ISLR Chap 7)

Assumptions on Errors

(ISLR Section 3.3.3)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \ldots, n$$

Assumptions on Errors

(ISLR Section 3.3.3)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \ldots, n$$

The least squares solution can be uniquely computed under the assumption:

- $X^\top X$ is invertible

Assumptions on Errors

(ISLR Section 3.3.3)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \ldots, n$$

The least squares solution can be uniquely computed under the assumption:

- $X^\top X$ is invertible

Least squares approach works well under the following assumptions:

- Var(ϵ_i) is the same
- $\epsilon_1, \ldots, \epsilon_n$ uncorrelated
- (n large, p small, $X^T X$ far from singular, etc.)

Assumptions on Errors

(ISLR Section 3.3.3)

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \ldots, n$$

The least squares solution can be uniquely computed under the assumption:

- $X^\top X$ is invertible

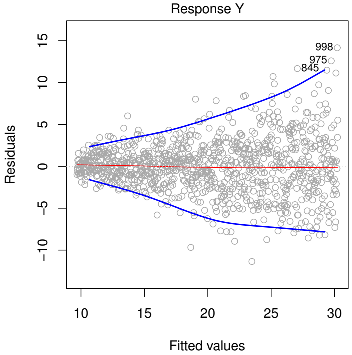
Least squares approach works well under the following assumptions:

- Var(ϵ_i) is the same
- $\epsilon_1, \ldots, \epsilon_n$ uncorrelated
- (n large, p small, $X^T X$ far from singular, etc.)

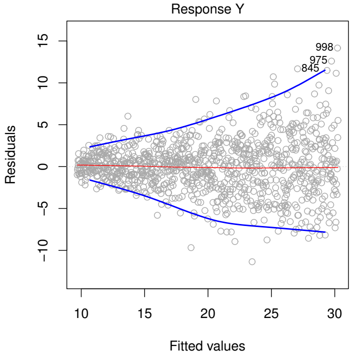
Some model evaluation techniques (p-values, conf. int.) are based on the following *additional* assumptions:

- ϵ_i is (approximately) mean-zero Gaussian.

Non-constant Error Variance (Heteroscedasticity)



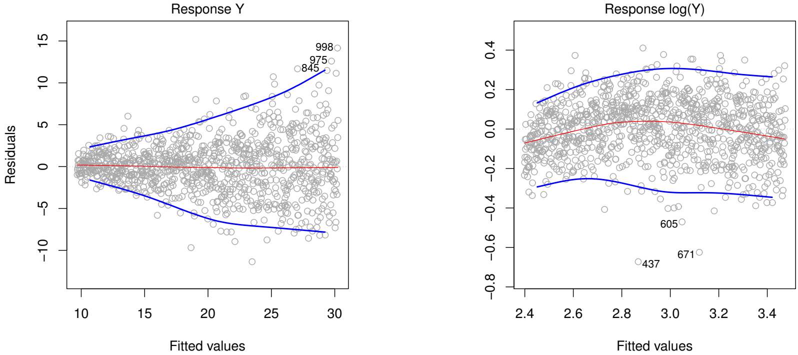
Non-constant Error Variance (Heteroscedasticity)



Possible solutions:

- Transformation of response. E.g. $\log Y$ or $\sqrt{Y}$

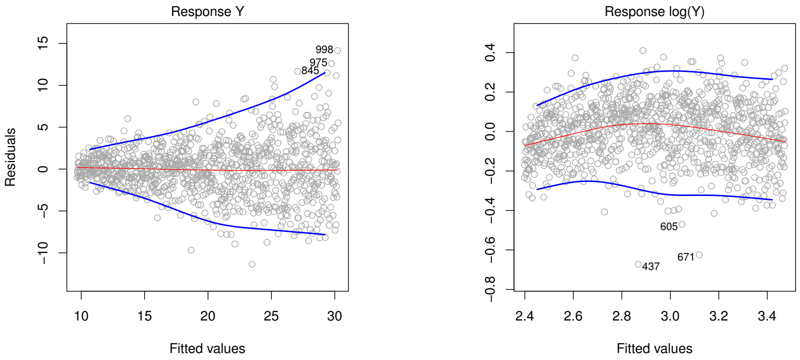
Non-constant Error Variance (Heteroscedasticity)



Possible solutions:

- Transformation of response. E.g. $\log Y$ or $\sqrt{Y}$

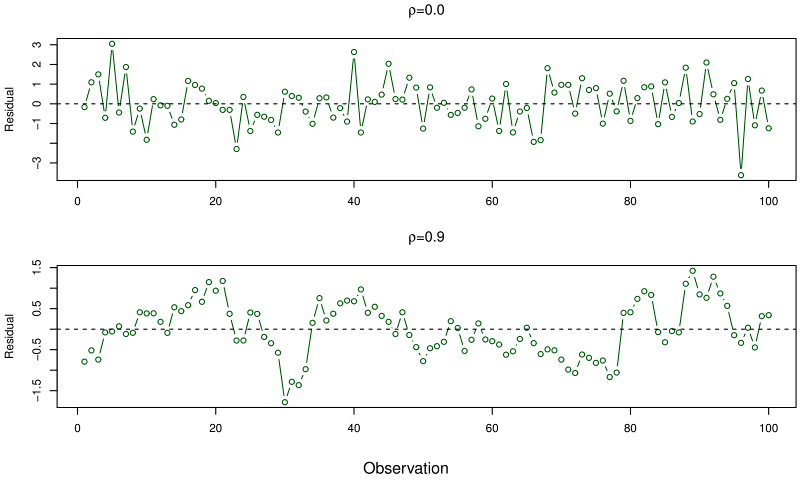
Non-constant Error Variance (Heteroscedasticity)



Possible solutions:

- Transformation of response. E.g. $\log Y$ or $\sqrt{Y}$
- Weighted least squares: weight $\propto \frac{1}{\mathrm{Var}(\epsilon_i)}$ (if $\mathrm{Var}(\epsilon_i)$ are known)

Correlated Errors



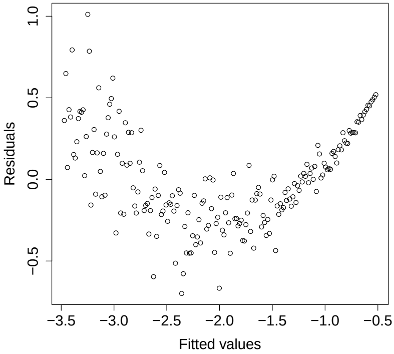
The residual plot on the right indicates that

A The errors have non-constant variance but the linear assumption is correct.

B The errors have non-constant variance and the linear assumption is wrong.

C The linear assumption is correct and the errors have constant variance.

D The linear assumption is wrong and the errors have constant variance.



Outliers

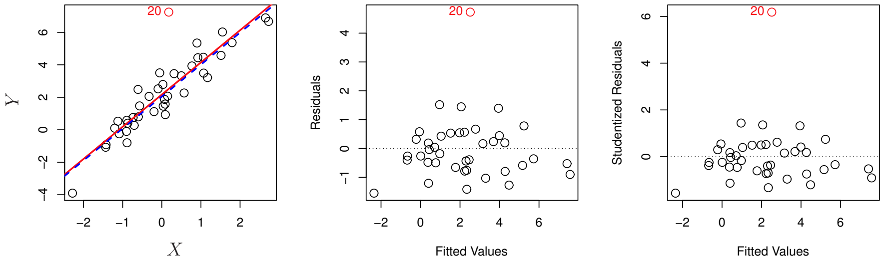
(ISLR Section 3.3.4)

A point $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$ with abnormal Y values

Outliers

(ISLR Section 3.3.4)

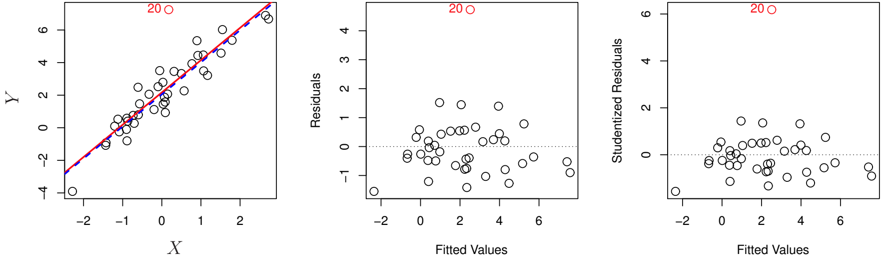
A point $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$ with abnormal Y values



Outliers

(ISLR Section 3.3.4)

A point $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$ with abnormal Y values



Detecting outliers: studentized residuals

High Leverage Points

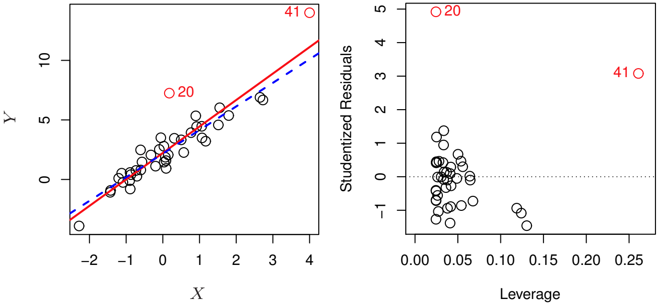
(ISLR Section 3.3.5)

A point $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$ with abnormal X values

High Leverage Points

(ISLR Section 3.3.5)

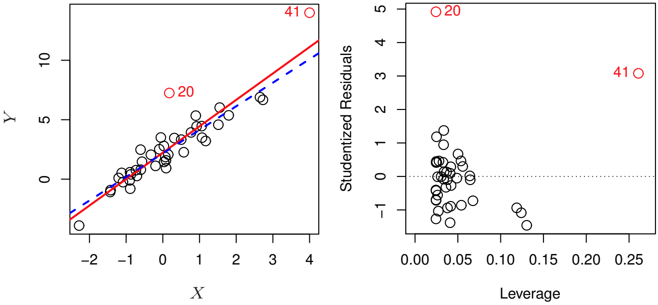
A point $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$ with abnormal X values



High Leverage Points

(ISLR Section 3.3.5)

A point $(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)$ with abnormal X values



Detecting high leverage points: leverage statistic

Summary: Other considerations in regression

Summary: Other considerations in regression

- Categorical/qualitative predictors:
Pick a base type; encode a k-level predictor with $k-1$ dummy variables

Summary: Other considerations in regression

- Categorical/qualitative predictors:
  Pick a base type; encode a k-level predictor with $k - 1$ dummy variables

- Nonlinearity:
  Detect by residual plots
  Add terms: X_1X_2, X_1^3, $\sqrt{X_1}$, $\log X_1$. Don't overfit!

Summary: Other considerations in regression

- Categorical/qualitative predictors:
Pick a base type; encode a k-level predictor with $k-1$ dummy variables

- Nonlinearity:
Detect by residual plots
Add terms: X_1X_2, X_1^3, $\sqrt{X_1}$, $\log X_1$. Don't overfit!

- Error assumptions:
– Non-constant variance: residual plots; transform response, weighted LS
– Correlated errors: look at residuals

Summary: Other considerations in regression

- Categorical/qualitative predictors:
Pick a base type; encode a k-level predictor with $k-1$ dummy variables

- Nonlinearity:
Detect by residual plots
Add terms: X_1X_2, X_1^3, $\sqrt{X_1}$, $\log X_1$. Don't overfit!

- Error assumptions:
– Non-constant variance: residual plots; transform response, weighted LS
– Correlated errors: look at residuals

- Outliers: Detect by studentized residuals and leverage statistics

Some of the figures in this presentation are taken from “An Introduction to Statistical Learning, with applications in R” (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Some of the figures in this presentation are created by Igor Labutov.