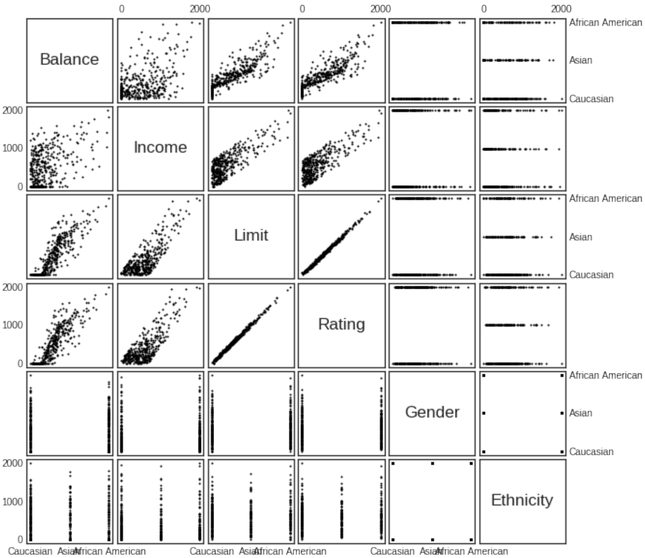
Linear Regression 3

Credit balance data



Predictor with two levels

Find the difference in credit card balance (y_i) between **male** and **female** (x_i).

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

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Predictor with two levels

Find the difference in credit card balance (y_i) between **male** and **female** (x_i).

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

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

Estimates of coefficients

	$\hat{\beta}_i$	SE($\hat{\beta}_i$)	t-statistic	p-value
Intercept	509.80	33.13	15.389	<0.0001
gender(Female)	19.73	46.05	0.429	0.6690

$$\hat{y}_i = 509.80 + 19.73x_i.$$

Main takeaway:

1. Male has credit card debt of 509.80 **on average**.
2. Female has credit card debt of 509.80+19.73 = 529.53 **on average**.
3. The difference in credit card debt is $\hat{\beta}_1 = 19.73$ **on average**.

Estimates of coefficients

	$\hat{\beta}_i$	SE($\hat{\beta}_i$)	t-statistic	p-value
Intercept	509.80	33.13	15.389	$<$0.0001
gender(Female)	19.73	46.05	0.429	0.6690

$$\hat{y}_i = 509.80 + 19.73x_i.$$

Main takeaway:

1. Male has credit card debt of 509.80 **on average**.
2. Female has credit card debt of 509.80+19.73 = 529.53 **on average**.
3. The difference in credit card debt is $\hat{\beta}_1 = 19.73$ **on average**.

Question: Can we conclude that females have more credit debt on average than males?

Predictor with more than two levels

Find the difference in credit card balance (y_i) between **Asian**, **Caucasian** and **Black**.

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

Predictor with more than two levels

Create two **dummy variables** x_{i1} and x_{i2} :

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

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

Using x_{i1} and x_{i2}, the regression can be written as

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

Estimates of coefficients

	$\hat{\beta}_i$	SE($\hat{\beta}_i$)	t-statistic	p-value
Intercept	531.00	46.32	11.464	<0.0001
ethnicity (Asian)	-18.69	65.02	-0.287	0.7740
ethnicity (Caucasian)	-12.50	56.68	-0.221	0.8260

Main takeaway: **On average,**

1. Black has credit debt of 531.00 .
2. Asian has 18.69 less debt than the Black.
3. Caucasian has 12.50 less debt than the Black.
4. Asian has ________________ less debt than Caucasian.

Estimates of coefficients

	$\hat{\beta}_i$	SE($\hat{\beta}_i$)	t-statistic	p-value
Intercept	531.00	46.32	11.464	<0.0001
ethnicity (Asian)	-18.69	65.02	-0.287	0.7740
ethnicity (Caucasian)	-12.50	56.68	-0.221	0.8260

Main takeaway: **On average,**

1. Black has credit debt of 531.00 .
2. Asian has 18.69 less debt than the Black.
3. Caucasian has 12.50 less debt than the Black.
4. Asian has ______________ less debt than Caucasian.

Question: How can we decide if there is any difference in credit card balance between the ethnicities?

Linear model diagnosis

Our model

Recall the linear regression model on n data points:

$$
\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \ldots + \epsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \ldots + \epsilon_2 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \ldots + \epsilon_n
\end{aligned}
$$

In this model, we assume that

Our model

Recall the linear regression model on n data points:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \ldots + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \ldots + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \ldots + \epsilon_n \end{aligned}$$

In this model, we assume that

1. $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are **independent**.

Our model

Recall the linear regression model on n data points:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \ldots + \epsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \ldots + \epsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \ldots + \epsilon_n \end{aligned}$$

In this model, we assume that

1. $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are **independent**.
2. $\epsilon_i \sim N(0, \sigma^2)$. Specifically, **they share the same variance** σ^2.

1. If $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are **not** independent

   Then, all tests in the previous lecture are invalid:

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

$$\frac{\hat{\beta}_i}{\text{SE}(\hat{\beta})}$$

2. If $\epsilon_1, \ldots, \epsilon_n$ do **not** share the same variance

   There is no closed-form formula for Cov($\hat{\boldsymbol{\beta}}$)

But how can we check these assumptions?

But how can we check these assumptions?

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

But how can we check these assumptions?

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \beta_p x_{ip}$$

But how can we check these assumptions?

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i \\
\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \beta_p x_{ip} \\
&\approx \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}
\end{aligned}
$$

But how can we check these assumptions?

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i \\
\hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \ldots + \beta_p x_{ip} \\
&\approx \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}
\end{aligned}
$$

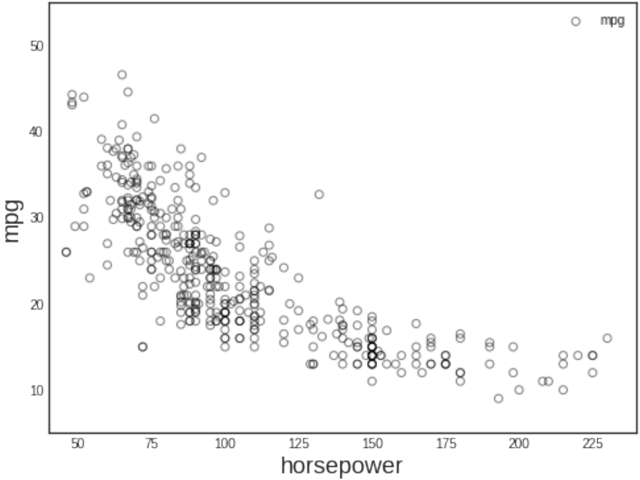
We will check if the **residuals**:

$$\text{residual of the i-th point} = y_i - \hat{y}_i \approx \epsilon_i$$

satisfy all these assumptions

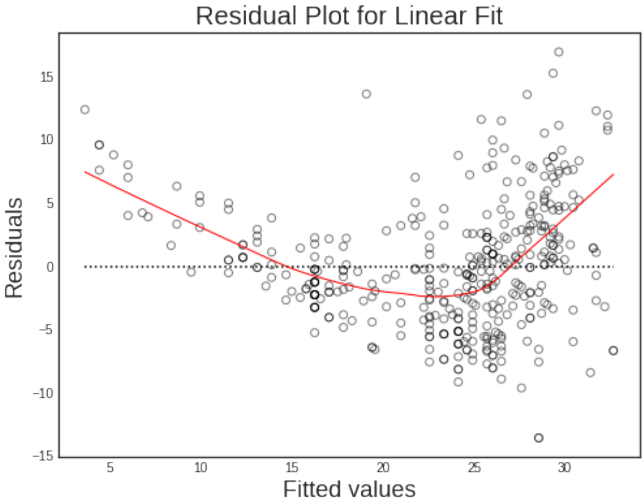
1. Non-linearity of the data

- Maybe the relationship between the predictors and the response is non-linear.



Residual plot

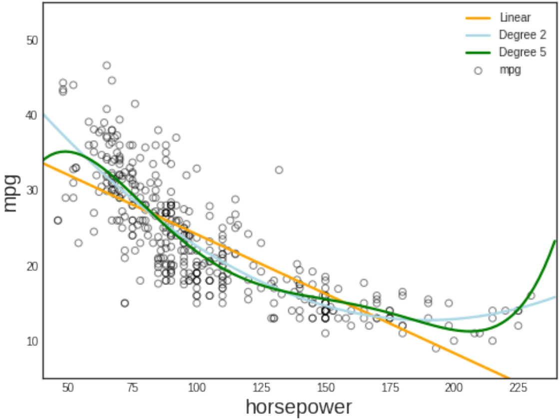
- Plot between the **fitted values** $\hat{y}_i$ and the **residuals** $y_i - \hat{y}_i$.



Non-linear regression

Try a polynomial function of the horsepower:

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



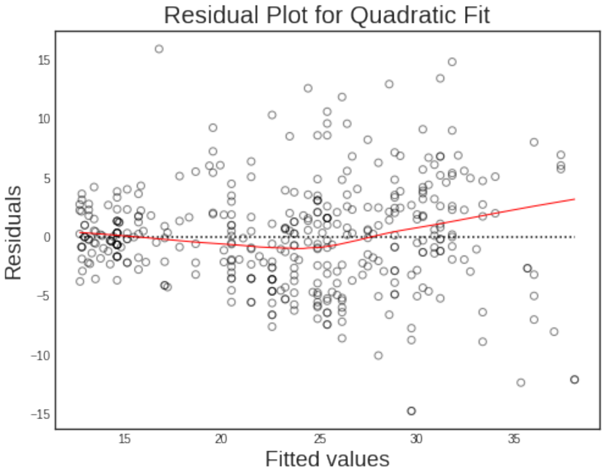
Estimates of coefficients

	$\hat{\beta}_i$	SE($\hat{\beta}_i$)	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	$<$0.0001
horsepower	-0.4662	0.0311	-15.0	$<$0.0001
horsepower^2	-0.0012	0.0001	10.1	$<$0.0001

Two things indicate that the quadratic fit is better:

- The p-value of `horsepower`^{2} is significant.
- The R^2 of this model is 0.688 compared to 0.606 of the linear model.

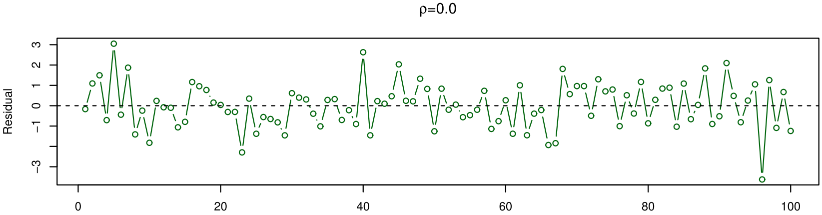
Residual plot of non-linear regression



The pattern disappears

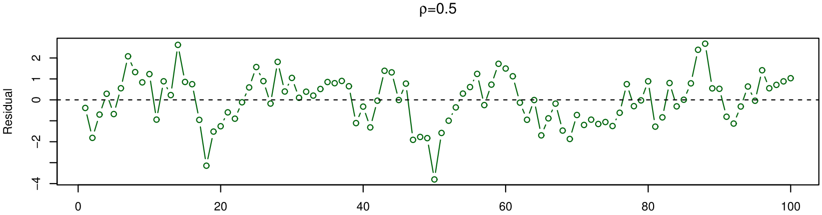
2. Correlation of error terms

$$\rho = \text{corr}(\epsilon_{i-1}, \epsilon_i)$$



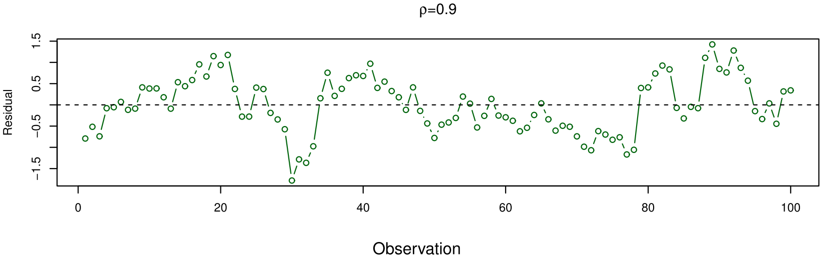
2. Correlation of error terms

$$\rho = \text{corr}(\epsilon_{i-1}, \epsilon_i)$$



2. Correlation of error terms

$$\rho = \text{corr}(\epsilon_{i-1}, \epsilon_i)$$



Durbin-Watson test

used to test if there is any correlation in the error terms

H_0 :There is no correlation among the residuals

H_1 :The residuals are autocorrelated

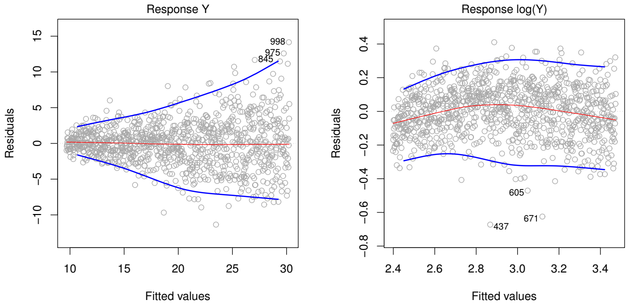
The test statistic is

$$d = \sum_{i=2}^{n} (e_i - e_{i-1})^2 \, / \sum_{i=1}^{n} e_i^2$$

Procedure: Choose a significance level α, then look up the value of d_L and d_U

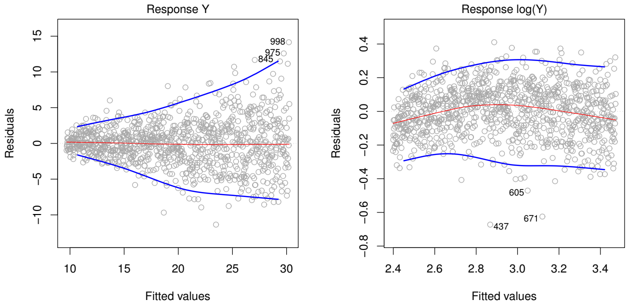
- Reject H_0 if $d < d_L$
- Do not reject H_0 if $d > d_U$
- Test inconclusive if $d_L < d < d_U$

3. Non-constant variance of error terms



- The variance increases as the fitted value increases.

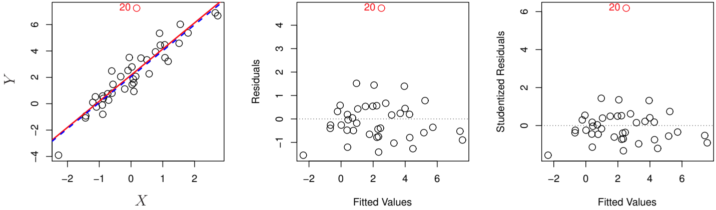
3. Non-constant variance of error terms



- The variance increases as the fitted value increases.
- Try transformation $Y \to \log(Y)$ or $Y \to \sqrt{Y}$ before fitting the model.

4. Outliers

A single point can heavily influence the RSE and R^2 of the model.

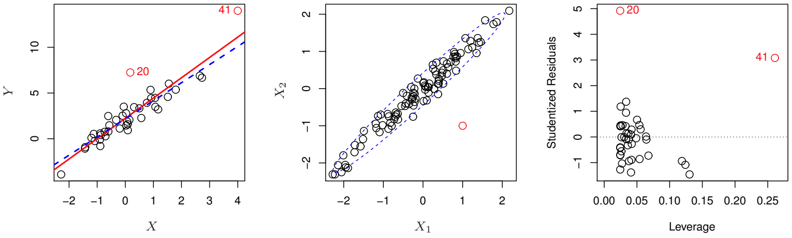


	RSE	R^2
Model with outlier	1.09	0.805
Model without outlier	0.77	0.892
Improvement	29%	11%

5. High leverage points

- **High leverage point** is a point with an unusual value of x_i.

- Detect high leverage points using the **leverage statistic**.



6. Collinearity

- **collinearity problem** happens when two predictors are highly correlated to each other.

- Highly correlated variables cause problems when fitting the model.

6. Collinearity

Example: Suppose we have data (x_i, y_i, z_i) from the true model:

$$y = 2x + 3z + \epsilon$$

and assume that $z = x$.

6. Collinearity

Example: Suppose we have data (x_i, y_i, z_i) from the true model:

$$y = 2x + 3z + \epsilon$$

and assume that $z = x$.
Fitted model can be:

$$\hat{y}_i = 2x_i + 3z_i$$

6. Collinearity

Example: Suppose we have data (x_i, y_i, z_i) from the true model:

$$y = 2x + 3z + \epsilon$$

and assume that $z = x$.
Fitted model can be:

$$\hat{y}_i = 2x_i + 3z_i$$

$$\text{or } \hat{y}_i = 1x_i + 4z_i$$

6. Collinearity

Example: Suppose we have data (x_i, y_i, z_i) from the true model:

$$y = 2x + 3z + \epsilon$$

and assume that $z = x$.
Fitted model can be:

$$\begin{aligned} \hat{y}_i &= 2x_i + 3z_i \\ \text{or } \hat{y}_i &= 1x_i + 4z_i \\ \hat{y}_i &= 5x_i \end{aligned}$$

6. Collinearity

Example: Suppose we have data (x_i, y_i, z_i) from the true model:

$$y = 2x + 3z + \epsilon$$

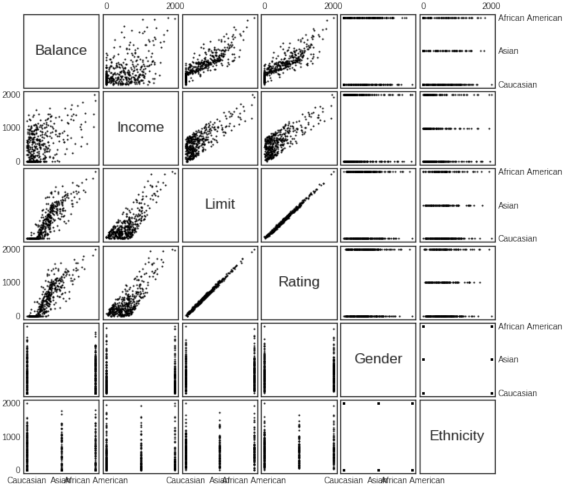
and assume that $z = x$.
Fitted model can be:

$$\begin{aligned} \hat{y}_i &= 2x_i + 3z_i \\ \text{or } \hat{y}_i &= 1x_i + 4z_i \\ \hat{y}_i &= 5x_i \end{aligned}$$

Fitting algorithm does not know which is the true model!

Credit balance data

Detect collinearity using **correlation matrix**. Remove a variable if the correlation is close to -1 or 1.



Multicollinearity

Multicollinearity happens when a predictor is a linear combination of other predictors.

Multicollinearity

Multicollinearity happens when a predictor is a linear combination of other predictors.

Example: Predictors x_i, z_i and w_i where $x_i = z_i + 2w_i$.

Multicollinearity

Multicollinearity happens when a predictor is a linear combination of other predictors.

Example: Predictors x_i, z_i and w_i where $x_i = z_i + 2w_i$.

Cannot be detected with correlation matrix. Instead, we use **variance inflation factor**

$$VIF(\hat{\beta}_i) = \frac{1}{1 - R^2_{X_i|X_{-i}}},$$

where $R^2_{X_i|X_{-i}}$ is the R^2 from a regression of X_i onto all other predictors.

Variance inflation factor

$$VIF(\hat{\beta}_i) = \frac{1}{1 - R^2_{X_i|X_{-i}}}.$$

[High multicol. in X_i] $\rightarrow$ [$R^2_{X_i|X_{-i}}$ is close to 1] $\rightarrow$ [high $VIF(\hat{\beta}_i)$]

Variance inflation factor

$$VIF(\hat{\beta}_i) = \frac{1}{1 - R^2_{X_i|X_{-i}}}.$$

[High multicol. in X_i] $\rightarrow$ [$R^2_{X_i|X_{-i}}$ is close to 1] $\rightarrow$ [high $VIF(\hat{\beta}_i)$]

General rule: There is multicollinearity if VIF is higher than 5 or 10

Solution: Drop the variable (in this case, X_i).

Reference

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