# MLR: Potential Problems

DS 301

Iowa State University

## Assumptions for linear regression

1. Relationship between $Y$ and $X = (X_1, X_2, \ldots, X_p)$ is approximately linear.
2. $E(\epsilon) = 0$.
3. $\text{Var}(\epsilon) = \sigma^2$.
4. $\epsilon$'s are uncorrelated.

# When do each of the assumptions kick in?

$Y = f(X) + \varepsilon$

$\hookrightarrow f(X)?$

(1) $\Rightarrow f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$

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

$\hookrightarrow Y$ is random

so what we want to estimate is $f(X)$ (or $E(Y)$)

$E(Y) = \underline{\beta_0} + \underline{\beta_1} X_1 + \cdots + \underline{\beta_p} X_p$ (2)

$$\min(RSS) = \min_{\beta_0, \beta_1, \ldots \beta_p} \sum_{i=1}^{n} \underbrace{(y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots \hat{\beta}_p X_{ip}))^2}_{(4)}$$

(3) $\hat{\sigma}^2 \rightarrow$ inference $= \dfrac{\sum_{i=1}^{n} e_i^2}{n-(p+1)}$

$\left\{ \begin{array}{l} se(\hat{\beta}) \\ se(\hat{Y}) \\ se(\text{pred}) \end{array} \right.$

# Non-constant variance of error terms

We assume that the error terms have a constant variance:

$$\mathrm{Var}(\epsilon_i) = \sigma^2.$$

$\Rightarrow \mathrm{var}(Y_i) = \sigma^2$ : $\sigma^2$ is unknown, we estimate it from data

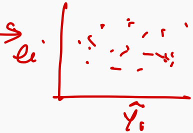
- The standard errors of our estimates rely on this assumption.
- Additionally, carrying out hypothesis tests, constructing prediction intervals, and confidence intervals associated with the linear model also rely upon this assumption.

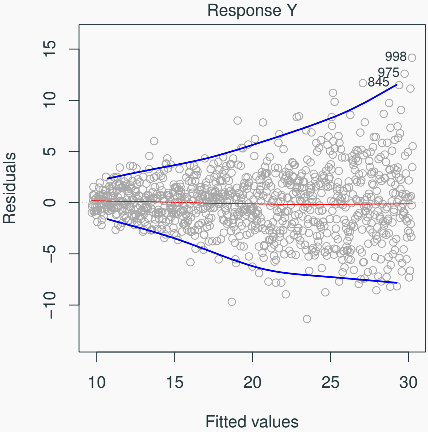
# Non-constant variance of error terms

- It may be the case that the variances of the error terms are non-constant.
- For example, the variances of the error terms may increase with the value of the response.
- How might we identify whether or not this is a problem with our model? Does this constant variance assumption hold?

# Residual plot

To diagnose this, we can plot residuals ($e_i$) vs. fitted values ($\hat{y}_i$) from our model. **If the constant variance assumption holds**, your plot should exhibit random scatter (no discernible pattern). If you see a funnel shape, there is a problem.

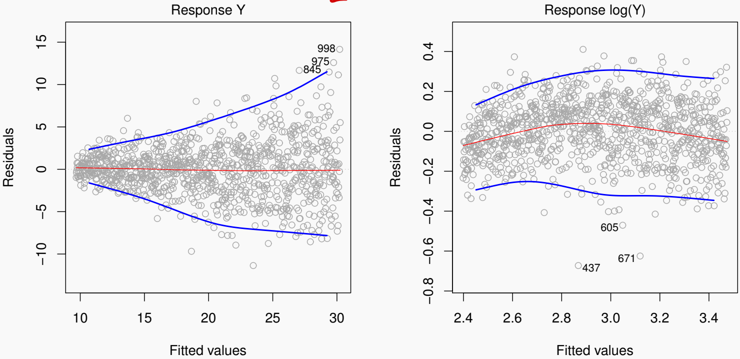


# Non-constant variance of error terms

One possible solution: transform the response $Y$ using a concave function such as $\log Y$ or $\sqrt{Y}$.

lm ( log (Y) ~ X ) .

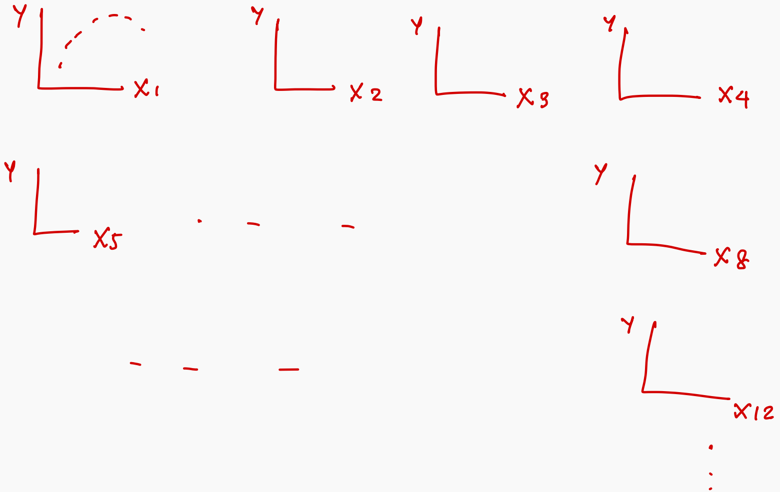


# Non-linearity of the data

- The linear regression model assumes that there is a straight-line relationship between the predictors and the response.
- If the true relationship is far from linear, then virtually all of the conclusions that we draw from the model are suspect.
- Additionally, the prediction accuracy of the model can be significantly reduced.

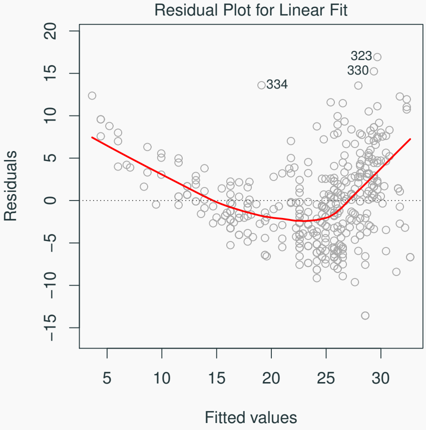
# How to diagnose non-linearity when you have multiple predictors?

$Y \quad X_1, \quad X_2, \ldots, X_{20}.$



# Residual plots: $e_i$ verus $\hat{y}_i$

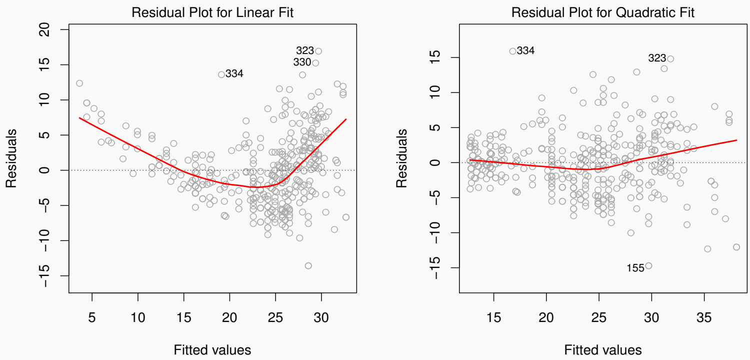
Ideally, the residual plot will show no discernible pattern. The presence of a pattern may indicate a problem with some aspect of the linear model.




ideal: random scatter.

# Simple solution : $\Rightarrow$ transform $X$

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, such as $\log(X)$, $\sqrt{X}$, and $X^2$, in the regression model.



# Polynomial Regression

$Y = B_0 + B_1 X_1 + B_2 X_2 + \mathcal{E}$

$Y = B_0 + B_1 X_1 + B_2 X_1^2 + B_3 X_2 + \mathcal{E}$

$Y = B_0 + B_1 X_1 + B_2 X_1^2 + B_3 X_1^3 + B_4 X_2 + \mathcal{E}$.

$$Y_i = \beta_0 + \beta_1 \underline{X_i} + \beta_2 X_i^2 + \beta_3 X_i^3 + \ldots + \beta_d X_i^{d} + \epsilon_i.$$

The coefficients here can be easily estimated using least squares because this is **still considered a standard linear model**.

$Y = B_0 + B_1 X_1 + B_2 Z_2 + B_3 Z_3 + B_4 X_2 + \mathcal{E}$.

Importantly, this means that all the inference tools for linear models (standard errors, F-tests, etc.) are all available in this setting.

# Dealing with non-linear relationships

- This process depends heavily on insight from exploratory data analysis. No shortcuts here.
- 'Linear' regression models actually includes a huge range of models.
  - Transform $Y$. (non. constant variance)
  - Transform predictors $X$. (linearity problem)
  - Polynomial regression.
  - Other models: piecewise polynomial regression, regression splines.

# Example

See R script: `MLR_Transformations.R`

# Multicollinearity

when you have predictors that
are correlated, you may
observe this phenomenon
( sig. F. test,
non-sig t. tests)

```
> summary(lm1)

Call:
lm(formula = y ~ X1 + X2 + X3)

Residuals:
     Min       1Q   Median       3Q      Max
-17.4784  -5.9323  -0.3146   5.9889  19.3380

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.3611     0.8718  -0.414    0.680
X1            0.6551     2.0999   0.312    0.756
X2            2.5562     2.3803   1.074    0.286
X3            3.5838     2.2600   1.586    0.116

Residual standard error: 8.65 on 96 degrees of freedom
Multiple R-squared:  0.3806,	Adjusted R-squared:  0.3612
F-statistic: 19.66 on 3 and 96 DF,  p-value: 5.107e-10
```

# Multicollinearity

Y ~ limit + rating
cc balance

Refers to the situation when two or more predictors are highly correlated.

weight

$$Y \sim X_1 + X_2 + X_3$$

chol. level height bmi

- When two or more predictors are highly correlated, it makes it difficult to separate out individual effects of predictors on the response.
- Incorporating redundant information in your model.
- Given $X_1$ is in the model, $X_2$ is not helping to explain much of $Y$ (and vice versa).

# Consequences of multicollinearity

$Y \sim X_1 + X_2$, $X_1$ and $X_2$ are perfectly correlated.

$X_1 = a + X_2$ , a, b are constants

Data set:

| $Y$ | $X_1$ | $X_2$ |
| --- | --- | --- |
| 2 | 1 | 1 |
| 3 | 1.5 | 1.5 |
| 6 | 3 | 3 |

| $\hat{B}_0$ | $\hat{B}_1$ | $\hat{B}_2$ | RSS |
| --- | --- | --- | --- |
| 0 | 1 | 1 | 0 |
| 0 | 2 | 0 | ⋮ |
| 0 | 0 | 2 | ⋮ |
| ⋮ | | | |

least square estimates $\hat{B}_0, \hat{B}_1, \hat{B}_2$

that minimizes $\sum_{i=1}^{n} (y_i - (\hat{B}_0 + \hat{B}_1 X_{i1} + \hat{B}_2 X_{i2}))^2$

$$\left.\begin{aligned} = \; & (2 - (\hat{B}_0 + \hat{B}_1 (1) + \hat{B}_2 (1)))^2 \\ + \; & (3 - (\hat{B}_0 + \hat{B}_1 (1.5) + \hat{B}_2 (1.5)))^2 \\ + \; & (6 - (\hat{B}_0 + \hat{B}_1 (3) + \hat{B}_2 (3)))^2 \end{aligned}\right\}$$

# Consequences of multicollinearity

- When your predictors are **perfectly correlated**, there is no unique set of least square solutions.
- In real applications, it is more likely you will have predictors that are **highly correlated** (not necessarily perfectly correlated). In this case, we can still obtain unique least square solutions but there is a great deal of uncertainty in our estimates $\hat{\beta}$.
- That means the standard errors for our least square estimates could be very large. $se(\hat{B})$

# Consequences of multicollinearity

$X_1, X_2 \longrightarrow \hat{B}_1, \hat{B}_2$

- Reduces accuracy of $\hat{\beta}_j$ for those predictors $X_j$ that are correlated.
- Results in increased standard errors for those $\hat{\beta}_j$'s.
- Inference becomes problematic:

↳ hypothesis testing:

↳ wider CI, PI

$H_0: B_j = 0$ vs. $H_1: B_j \neq 0$.

$$ts = \frac{\hat{B}_j - B_j}{se(\hat{B}_j)}. \Rightarrow se(\hat{B}_j) \uparrow \text{ then } ts \downarrow$$

- we may fail to reject $H_0$ due to inflated $se(\hat{B}_j)$
- reduced power of test.

# How to detect multicollinearity among 2 or more predictors?

Variance Inflation Factor: VIF

- VIF $> 4$ or VIF $> 10$ may indicate a problem.

For implementation, see `R` script:
`example_multicollinearity.R`