Applied Regression II

CMED6020 – Session 4

Eric Lau (ehylau@hku.hk)

School of Public Health
The University of Hong Kong

1 Feb 2021

Session 4 learning objectives

After this session, students should be able to

- Apply poisson and negative binomial regression models to count data
- Identify and apply suitable model to overdispersed data
- Identify influential observations
- Perform model diagnostics
- Understand and deal with multicollinearity

Overdispersion

Recap of last session

- Poisson, linear and logistic regression all belong to the same family of generalized linear regression (GLM)
- We can model count data with Poisson regression

$$\log(E(Y)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$

- Interpretation is often expressed in terms of relative risk (RR) or incidence rate ratio (IRR), which is the exponential of β
- We can include an offset term to account for different exposure

Overdispersion

- In many situations, we find that count data have a particular property which makes the Poisson model unsuitable for use
- That is, in many applications we find that the variance of our count data is appreciably larger than the mean
  - (Poisson model assumes variance equals the mean: $Var(Y) = E(Y) = \mu$)
- This is known as ‘overdispersion’.
  - If we fit a Poisson model to overdispersed data, our results will not be reliable, so what can we do?

Generalising the Poisson model

- We can generalise the Poisson regression model to include a ‘random effect’ in the rate λ_i:

$$Y_i \sim \text{Poisson}(\mu_i), \text{ where } \mu_i = A_i\lambda_i \text{ and } \lambda_i = \exp(\beta' x_i + \varepsilon_i)$$

- Here ε_i is a ‘random effect’. We can think of ε_i as representing the effects of one or more unobserved explanatory variables

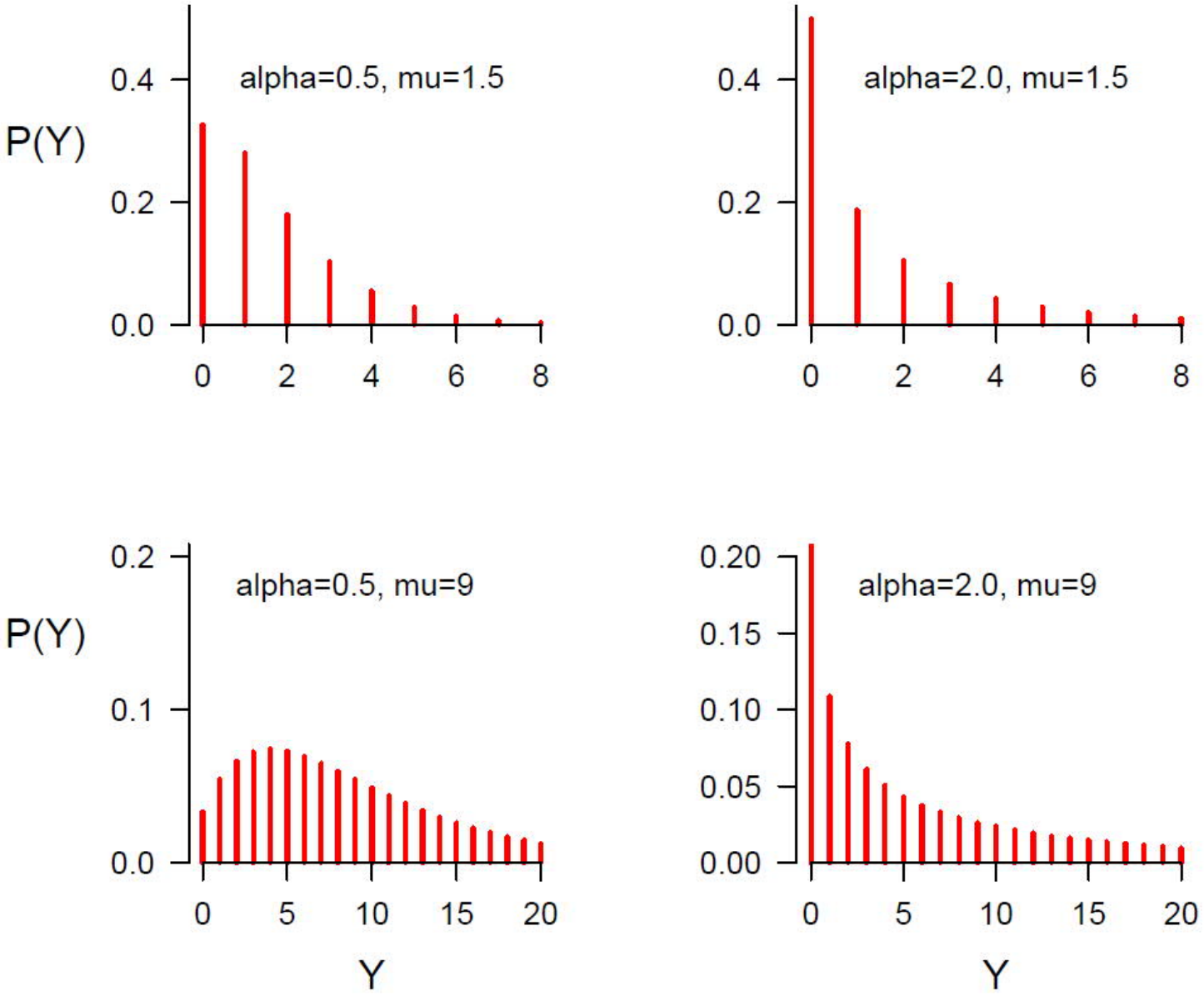
The Negative Binomial model

- In the special case where we allow $\exp(\varepsilon_i)$ to follow a gamma distribution with mean 1 and variance α, then Y_i follows a *negative binomial distribution* with
  - Mean = μ_i
  - Variance = $\mu_i + \alpha\mu_i^2$

- Check this link for pdf – http://en.wikipedia.org/wiki/Negative_binomial

```
#Example: lambda=5, gamma mean=1, variance=2
rpois(10000, 5*rgamma(10000,1/2,1/2))
```

Examples of NegBin distribution



Properties of the Negbin distribution

1. As μ increases, the 'mass' of the distribution shifts to the right. The mean value $E(Y) = \mu = A\lambda$
2. The variance is given by: $Var(Y) = \mu + \alpha\mu^2 > \mu$
3. As μ increases, the probability of 0 decreases.
4. As α decreases towards 0, the distribution becomes more like a Poisson

- Particularly note that α does not affect the mean
  - To check whether there is overdispersion in our data, we can fit the NegBin model and see how likely it is that $\alpha = 0$

Example – Epileptic seizures

- The file ‘exampleepilepsy.csv’ contains data from a RCT which compared progabide (therapy=1) with a placebo (therapy=0)
- For each patient we know:
  - the number of seizures (*x*) experienced in 8 weeks prior to randomisation;
  - the number of seizures (*y*) experienced during the 8-week trial (let's set $A = 1$ since all patients had the same follow-up)
  - We also know each patient's age
- Was progabide more effective than placebo in reducing seizure frequency?

Check the overdispersion

```
> mean(epilepsy$y)
[1] 28.41379
> var(epilepsy$y)
[1] 823.4749
```

We can see that y seems to be overdispersed, since the variance (823.475) is much larger than the mean (28.41)

Practice

Fit a poisson regression model for the number of seizures with an intercept only.

Does your model fit the data satisfactorily?

Fitting NegBin regression model in R

- Unfortunately, the glm function in R only fits negative binomial regression model when α is known.
- For the case when α is unknown and has to be estimated from the parameter, use glm.nb in the “MASS” (Modern Applied Statistics…) package

glm.nb(formula, data, subset, …)

- **formula, *data*** and ***subset*** same as glm
- ***family*** not needed

Fit a NegBin regression model

```
> require(MASS)
> summary(glm.nb(y~1, data=epilepsy))
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.3469     0.1141   29.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1.3898) family taken
   to be 1)
```

Hence the fitted mean (μ) is exp(3.347) = 28.4 (The observed mean was 28.4.)

Overdispersion

```
          Theta:  1.390
    Std. Err.:  0.253

  2 x log-likelihood:  -503.188
```

In glm.nb, θ = 1 / α, so the fitted variance ($\mu + \alpha\mu^2$) is $28.4 + 28.4^2 / 1.39 = 609.3$. (The observed variance was 823.5.)

Fit a NegBin regression model to estimate therapy effect

```
> nb.therapy1 <- glm.nb(y~therapy, data=epilepsy)
> round(exp(cbind(coef(nb.therapy1),
   confint(nb.therapy1))),3)
                      2.5 % 97.5 %
(Intercept) 34.393 25.522 47.760
therapy      0.664  0.428  1.028

> deviance(nb.therapy1)/df.residual(nb.therapy1)
[1] 1.14985
```

The estimated effect of therapy (unadjusted) is to reduce the rate of seizures by 34% (RR = 0.66, 95% CI = 0.43 -1.03)

The model fit the data satisfactorily.

Adjusting for the original covariates (*x*, age)

```
> nb.therapy2 <- glm.nb(y~therapy+x+age, data=epilepsy)
> round(exp(cbind(coef(nb.therapy2),
   confint(nb.therapy2))),3)
                   2.5 % 97.5 %
(Intercept) 6.485 2.865 14.747
therapy     0.841 0.620  1.141
x           1.031 1.023  1.039
age         1.016 0.991  1.042
```

Adjust for log-transfomed *x* and age

```
> epilepsy$log10x <- log10(epilepsy$x)
> nb.therapy3 <- glm.nb(y~therapy+log10x+age,
   data=epilepsy)
> round(exp(cbind(coef(nb.therapy3),
   confint(nb.therapy3))),3)
                    2.5 % 97.5 %
(Intercept) 0.922 0.345  2.462
therapy     0.737 0.553  0.981
log10x      8.813 5.592 13.989
age         1.014 0.990  1.038
```

The estimated effect of therapy (adjusted) is to reduce the rate of seizures by 26% (RR = 0.74, 95% CI = 0.55-0.98)

Comparing models

- We can compare models based on the Akaike Information Criterion (AIC), which is defined as

$$AIC = 2k - 2ln(L)$$

  where k is the number of parameters in the model, L is the likelihood
- A lower AIC indicates a ‘better’ model
- As a rule of thumb, a difference in AIC of 2 or more can be regarded as significant

Model checking

Model	terms	log-likelihood	k	AIC
0	Intercept	-251.6	2	507.2
1	therapy	-249.9	3	505.9
2	therapy age x	-226.4	5	462.9
3	therapy age $\log_{10} x$	-222.8	5	455.7

Note that in addition to the explanatory variables, each model includes an intercept term β_0 and the 'scale' parameter α. From the above table we conclude that the AIC favors Model 3.

```
AIC(nb.epilepsy0, nb.therapy1, nb.therapy2, nb.therapy3)

logLik(…)
```

Conclusions from epilepsy example data

Final model:

```
                  2.5 % 97.5 %
(Intercept) 0.922 0.345  2.462
therapy     0.737 0.553  0.981
log10x      8.813 5.592 13.989
age         1.014 0.990  1.038
> deviance(nb.therapy3)/df.residual(nb.therapy3)
[1] 1.165512
```

- We can present the adjusted relative risk of seizures for progabide compared to placebo, which is 0.74 (95% CI: 0.55-0.98)
- We can say that the seizure rate varies a great deal between patients
- We have seen that the log-transformed baseline seizure rate is a more useful explanatory variable than the original seizure rate
- The fit of the NegBin model seems to be acceptable

Practice

Fit a poisson regression model for the number of seizures, to estimate the therapy effect adjusted for $\log_{10}x$ and age.

Does your model fit the data satisfactorily?

How does it compare with the negative binomial model?

Association of gestational age and growth measures at birth with infection-related admissions to hospital throughout childhood: a population-based, data-linkage study from Western Australia

Jessica E Miller, Geoffrey C Hammond*, Tobias Strunk*, Hannah C Moore, Helen Leonard, Kim W Carter, Zulfiqar Bhutta, Fiona Stanley, Nicholas de Klerk, David P Burgner*

	Any infection	Invasive bacterial	Gastrointestinal	Lower respiratory tract
Gestational age (weeks)				
<28	2·91 (2·55–3·33)	2·63 (1·45–4·79)*	2·22 (1·60–3·09)	4·43 (3·66–5·36)
28–29	2·49 (2·24–2·78)	2·61 (1·65–4·13)	2·22 (1·76–2·80)	4·77 (4·07–5·59)
30–31	2·29 (2·07–2·53)	2·22 (1·49–3·33)	2·29 (1·90–2·75)	3·83 (3·32–4·43)
32–34	1·72 (1·64–1·81)	1·44 (1·15–1·81)*	1·64 (1·49–1·81)	2·48 (2·30–2·68)
35	1·57 (1·49–1·65)	1·72 (1·37–2·16)	1·64 (1·47–1·83)	1·91 (1·77–2·07)
36	1·44 (1·39–1·49)	1·28 (1·08–1·51)†	1·41 (1·32–1·51)	1·72 (1·61–1·84)
37	1·31 (1·28–1·34)	1·20 (1·07–1·35)	1·34 (1·28–1·40)	1·45 (1·38–1·52)
38	1·15 (1·13–1·17)	1·02 (0·94–1·12)§	1·13 (1·09–1·17)	1·22 (1·18–1·26)
39–40 (reference)	1·00	1·00	1·00	1·00
41	0·94 (0·92–0·96)	0·86 (0·78–0·95)†	0·92 (0·89–0·96)	0·96 (0·92–0·99)‡
≥42	0·99 (0·94–1·04)§	0·77 (0·60–1·00)‡	0·92 (0·83–1·01)§	1·08 (0·99–1·19)§

Table 2: Risk of childhood infection-related admissions to hospital

Statistical analysis

For each child included in the analysis, we calculated time at risk from birth-related hospital discharge to death, their 18th birthday, or end of the study (Dec 31, 2010), whichever occurred first. A χ^2 test of independence was done for children with and without recorded infection-related admissions to hospital with dichotomous and categorised measures. Likelihood ratio tests for linearity and departure from linearity were also done for ordered variables. The primary outcomes were the number and type of infection-related admissions to hospital. We calculated rate ratios (RR) for gestational age, birthweight, and birth length measures using a multilevel negative binomial regression framework (with births grouped by mother), and adjusted for maternal age at delivery (<20, 20–24, 25–29, 30–24, ≥35 years), birth year (2-year blocks), birth season, parity (previous

Miller et al., Lancet, 2016

Generalized linear model (GLM)

- Linear, logistic, poisson, negative binomial regression models all belong to GLM

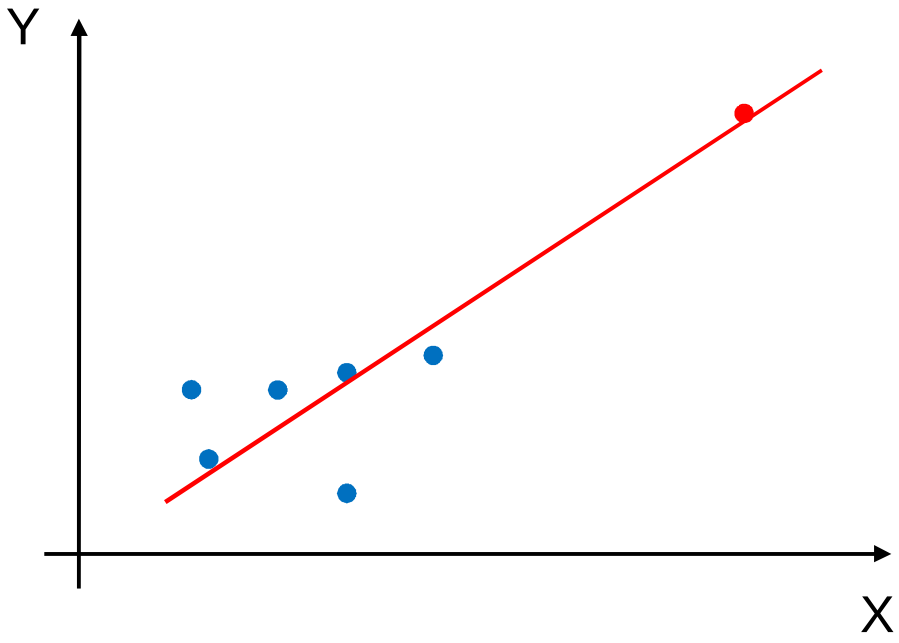
Main components of GLM:

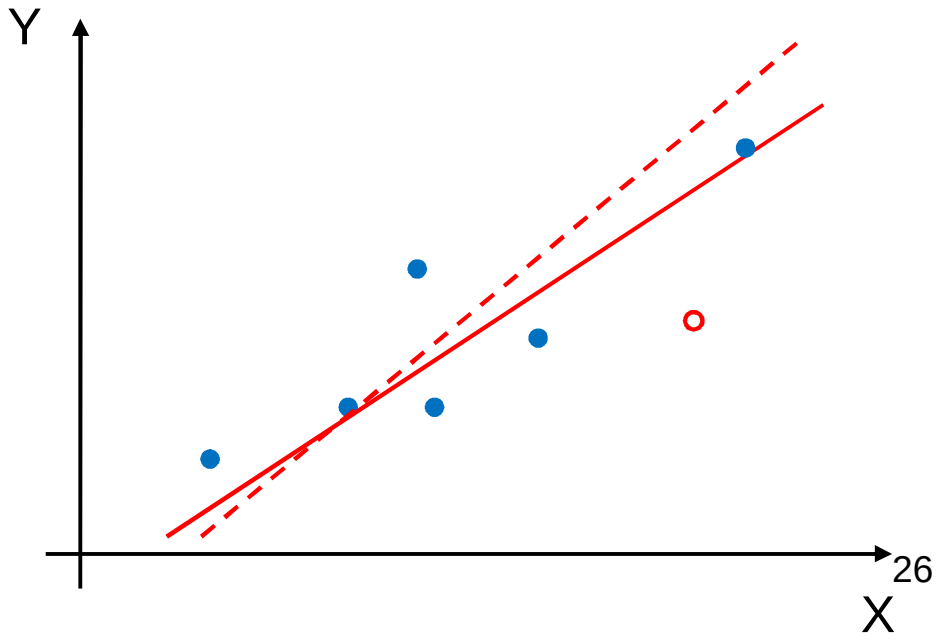
- Systematic component:
  - Link function *g* which links the outcome variable to the linear predictor
  - $g(E(Y)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots$
- Random component:
  - Specifying variance to mean relation
  - e.g. poisson: $\sigma^2 = \mu$
  - negative binomial: $\sigma^2 = \mu + \alpha\mu^2$

Influential observations

Influence of an observation

- Leverage
  - How much an observation differs from the mean of the predictor variables (i.e., unusual predictor values, left figure)
- Cook's distance
  - Change in predicted values if an observation were excluded (right figure)





Leverage

- Leverage values are often used for detecting observations that have a large impact on the predicted values
  - Due to unusual predictor values
- Bounded by 0 and 1
- Average value = k/n
  - k: number of estimated parameters in the model, including the constant/intercept
  - n: sample size
- Rough guideline: examine values greater than $2k/n$
  - In the MVC example ($MVC = \alpha + \beta_1 age + \beta_2 height$), $2k/n = 2 \times 3 / 41 = 0.1463$

MVC regression

```
> mvc <- read.csv("http://web.hku.hk/~ehylau/mvc.csv")
> mvc.lm <- lm(MVC ~ height + age, data=mvc)
> summary(mvc.lm)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -465.626    460.333  -1.011   0.3182
height         5.398      2.545   2.121   0.0405 *
age           -3.075      1.467  -2.096   0.0428 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.92 on 38 degrees of freedom
Multiple R-squared:  0.2612,    Adjusted R-squared:  0.2224
F-statistic: 6.719 on 2 and 38 DF,  p-value: 0.003173
```

MVC – revised dataset

- Suppose the height of obs #41 was 178cm (instead of 168cm)

```
> mvc.r <- mvc
> mvc.r$height[41] <- 178
> summary(lm(MVC ~ height + age, data=mvc.r))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -213.800    450.162  -0.475   0.6375
height         4.029      2.509   1.606   0.1166
age           -3.511      1.464  -2.397   0.0215 *
```

- Refit the model:
  - β_{age} = -3.0 → β_{age} = -3.5
  - due to only 1 single influential observation

Calculate leverage statistics in R

```
> mvc.r.lm <- lm(MVC ~ height + age, data=mvc.r)
> hatvalues(mvc.r.lm)
```

```
        1         2         3         4         5         6         7         8         9
0.13776073 0.07903710 0.07150184 0.07308259 0.05550758 0.05550758 0.15004733 0.05066261 0.07436060
       10        11        12        13        14        15        16        17        18
0.06144896 0.04614485 0.07723559 0.05573343 0.03832624 0.03038210 0.02846689 0.09832297 0.02948528
       19        20        21        22        23        24        25        26        27
0.02817681 0.03259448 0.05268320 0.05315284 0.02691674 0.07052102 0.05590691 0.05911810 0.06815930
       28        29        30        31        32        33        34        35        36
0.03800230 0.05860663 0.10987053 0.04925369 0.06024465 0.04560045 0.05036987 0.09734495 0.17414856
       37        38        39        40        41
0.11226514 0.10818219 0.14218370 0.11421299 0.17947069
```

```
> sort(hatvalues(mvc.r.lm), decreasing=T)
```

```
       41        36         7        39         1        40        37        30        38
0.17947069 0.17414856 0.15004733 0.14218370 0.13776073 0.11421299 0.11226514 0.10987053 0.10818219 …
```

Cook's distance

- A measure of the overall influence of a case on a model as a whole
- Tells us how much deleting a case affects not only the residual for that case, but also the residuals of the remaining cases
- Cook's distance depends on the standardized residuals of a case and its leverage.

$$D_i = \frac{Z_i^2 \times h_i}{(1 - h_i)k}$$

where Z_i is the standardized residual, h_i is the leverage,

k is the number of parameters

- Rough guideline: examine cases with $D_i > 4/n$
  - In the MVC example, $4/n = 4 / 41 = 0.0976$

Calculate Cook's distance in R

```
> round(cooks.distance(mvc.r.lm),3)

 1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16
0.055 0.026 0.005 0.000 0.101 0.012 0.037 0.088 0.044 0.001 0.002 0.002 0.049 0.037 0.000 0.001
   17    18    19    20    21    22    23    24    25    26    27    28    29    30    31    32
0.031 0.009 0.009 0.001 0.062 0.007 0.000 0.000 0.003 0.028 0.009 0.000 0.006 0.003 0.000 0.017
   33    34    35    36    37    38    39    40    41
0.049 0.001 0.001 0.004 0.077 0.000 0.001 0.045 0.350

> sort(round(cooks.distance(mvc.r.lm),2), decreasing=T)
41    5    8   37   21    1   13   33   40    7    9   14    2   17   26   32    6   18   19
0.35 0.10 0.09 0.08 0.06 0.05 0.05 0.05 0.05 0.04 0.04 0.04 0.03 0.03 0.03 0.02 0.01 0.01 0.01…
```

Identify influential observations automatically

```
> influence.measures(mvc.r.lm)
      dfb.1_  dfb.hght  dfb.age    dffit cov.r   cook.d    hat inf
1   0.260311 -0.215759 -0.34413  0.40550 1.157 5.48e-02 0.1378
2  -0.000466 -0.036551  0.21085 -0.27825 1.094 2.59e-02 0.0790
3  -0.019309  0.003888  0.09313 -0.11770 1.150 4.72e-03 0.0715
4  -0.000607  0.004762 -0.02322  0.03161 1.168 3.42e-04 0.0731
5  -0.127377  0.055068  0.43291 -0.58385 0.741 1.01e-01 0.0555   *
...
38 -0.011979  0.014757 -0.01836 -0.03109 1.214 3.31e-04 0.1082
39 -0.028874  0.033622 -0.02807 -0.05594 1.261 1.07e-03 0.1422   *
40  0.047084 -0.008476 -0.31767 -0.36876 1.124 4.53e-02 0.1142
41  0.737215 -0.653128 -0.91025 -1.08263 0.880 3.50e-01 0.1795   *
```

What to do with influential observations?

- Don’t throw away data easily without justification
- Perform sensitivity analysis
  - Do the results change qualitatively if we remove the outliers?
- Reassess if the data came from the target population

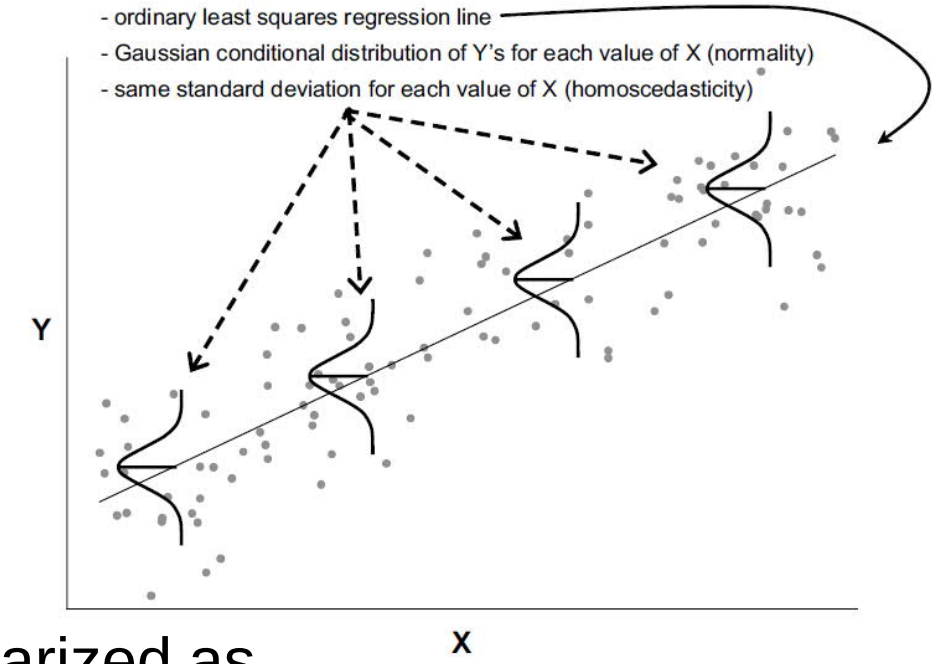
Model diagnostics

Why model diagnostics?

- Estimation method and statistical tests are based on model assumptions
- General consequences of violated assumptions
  - Biased estimated coefficients or standard errors
  - Inefficient estimation (achieving lower precision with the same sample size)
- Model diagnostics
  - Identify any potential violated assumptions
  - If yes, assess the extent of violation
  - Acknowledge limitation of the fitted model
  - Or suggest an alternative statistical model if assumptions seriously violated

Assumptions of linear regression model

- Linearity
  - Linear relationship between predictors and dependent variable
- Homoscedasticity
  - Constant variance of the errors
- Normality of the errors
- Independence
  - No correlation / autocorrelation between the errors



Marrie et al.,
J Clin Epidemiol, 2009

The last three assumptions can be summarized as

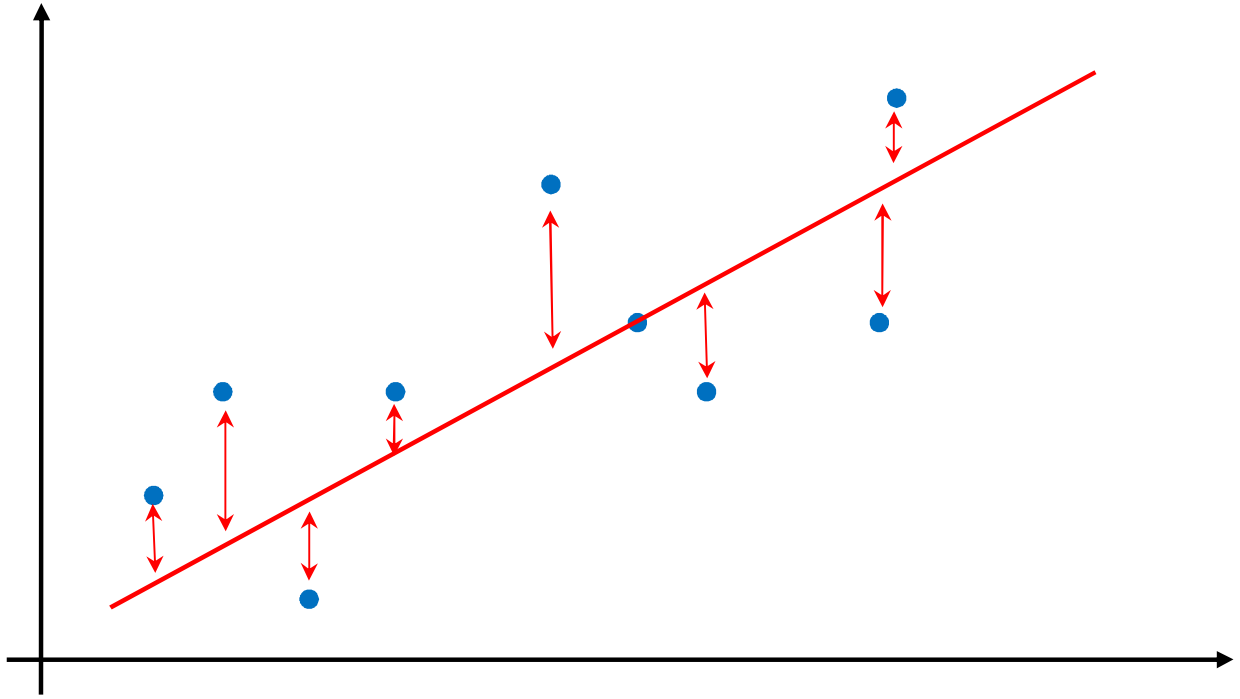
- ε iid ~ $N(0, \sigma^2)$
  - iid: independently and identically distributed
- Residual plot is very helpful to identify violation of assumptions

Residuals

- Definition:

$$e = y - \hat{y}$$

  - Difference between observed outcome and estimated value



Type of Residuals

- Raw residual:

$$e = y - \hat{y}$$

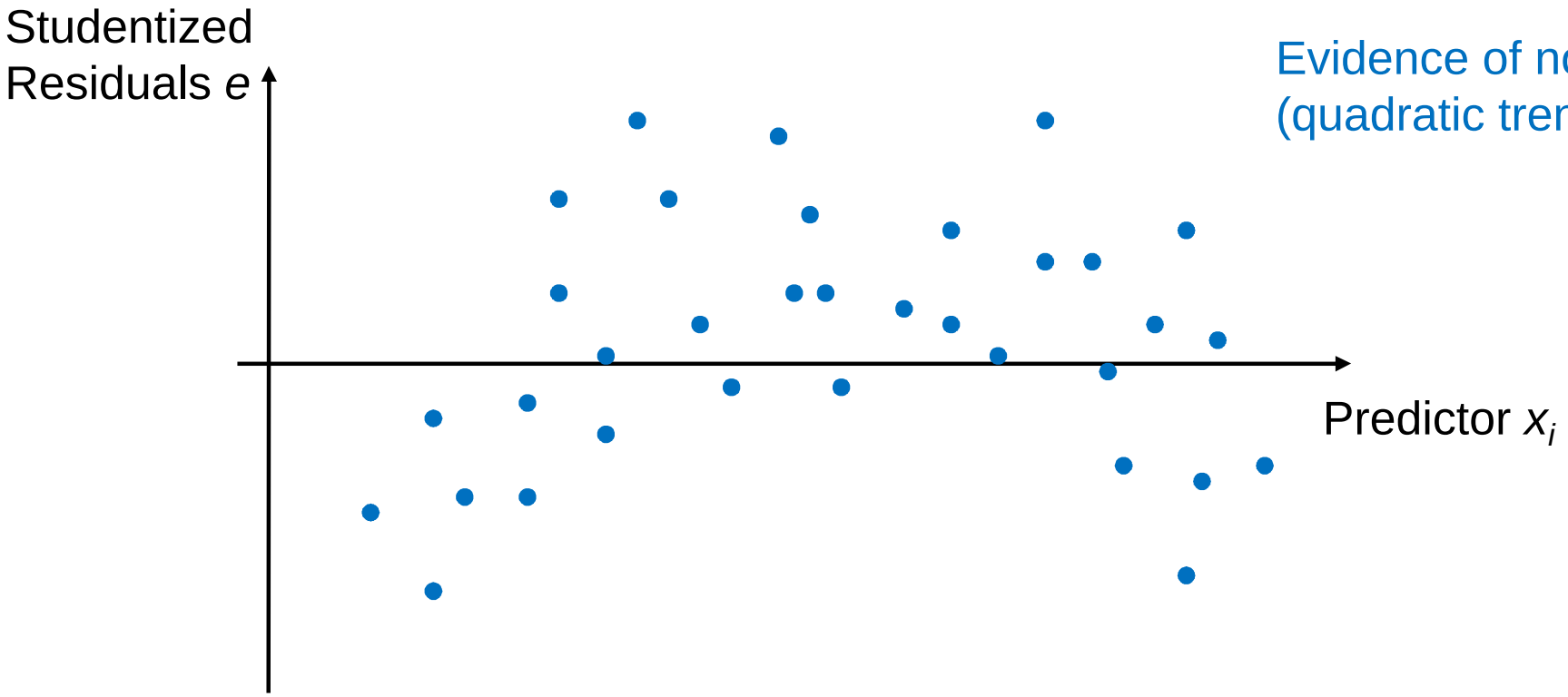
  - *model*$res in R
- Studentized residuals

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_i}}$$ (For reference only)

  - rstudent(*model*) in R
- Standardized residuals by each data point
- If model assumptions are correct, studentized residuals will
  - follow a t-distribution
  - mean = 0
  - variance = 1

Identify potential nonlinearity

- Plot studentized residuals against each predictor
  - Non-linear pattern suggests higher order terms or transformations of that predictor may be necessary
  - Less useful for categorical variables



MVC data - revisit

- Reading dataset

  mvc <- read.csv("http://web.hku.hk/~ehylau/mvc.csv")

- Fit a linear regression model for MVC on height and age:

  mvc.lm <- lm(MVC ~ height + age, data=mvc)

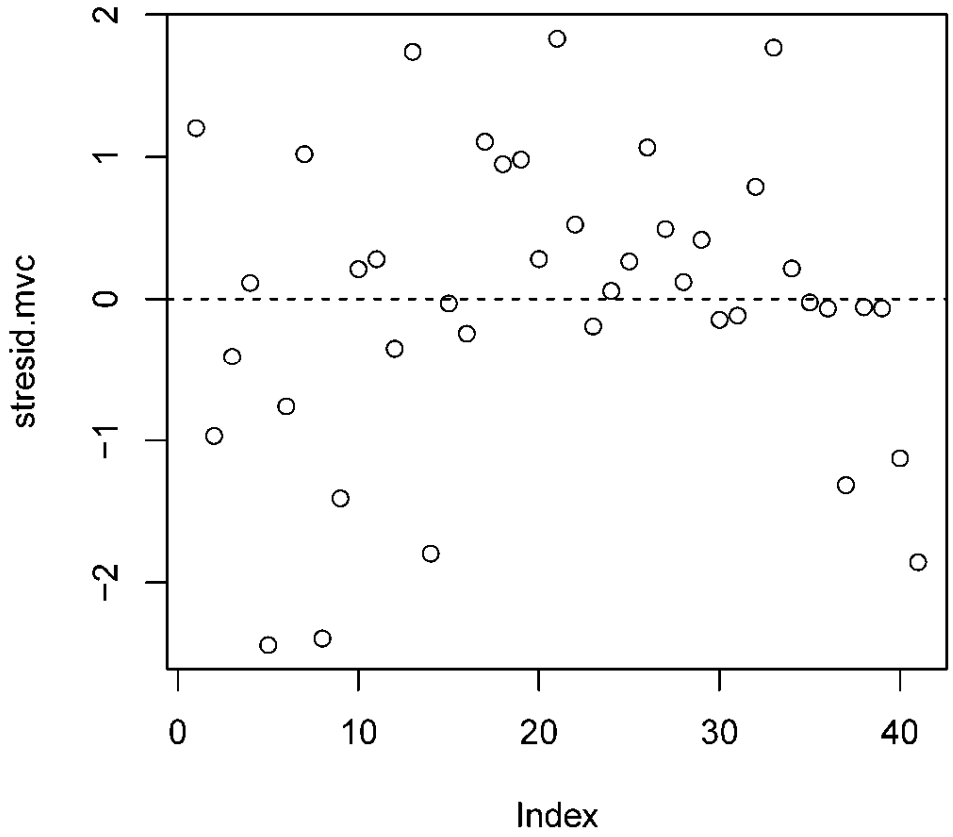
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -465.626    460.333  -1.011   0.3182
height         5.398      2.545   2.121   0.0405 *
age           -3.075      1.467  -2.096   0.0428 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.92 on 38 degrees of freedom
Multiple R-squared:  0.2612, Adjusted R-squared:  0.2224
F-statistic: 6.719 on 2 and 38 DF,  p-value: 0.003173
```


These results may not be reliable if the model assumptions were invalid

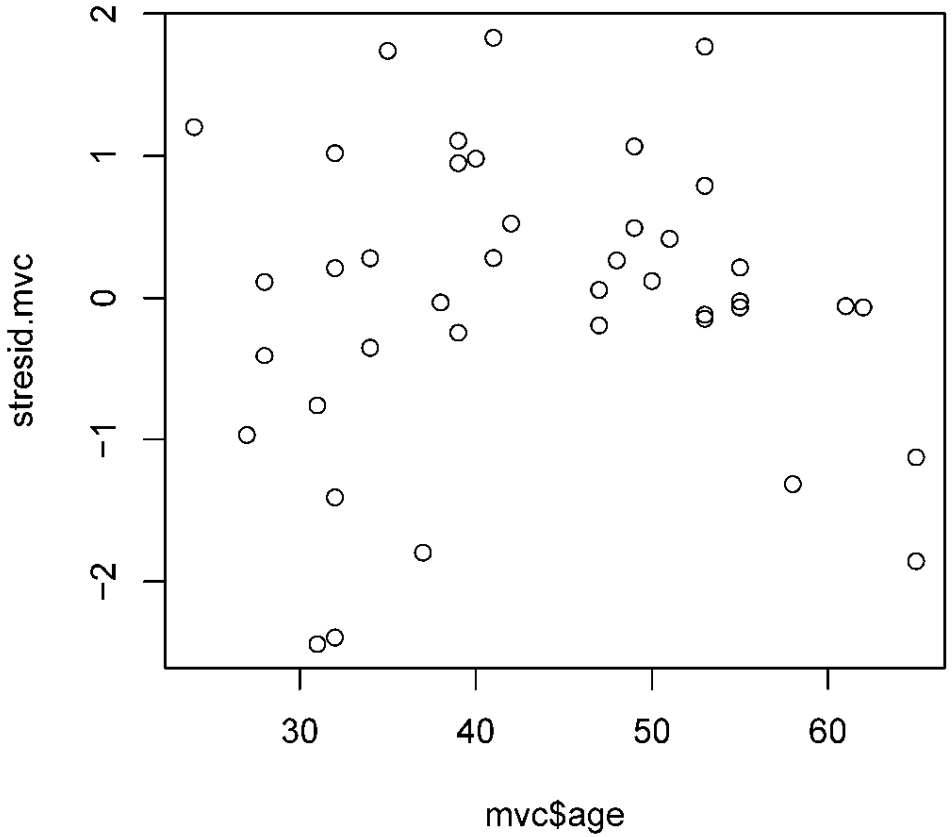
Create residuals and residual plot

```
> stresid.mvc <- rstudent(mvc.lm)
> plot(stresid.mvc)
```



MVC Residual plots – against predictors (age)

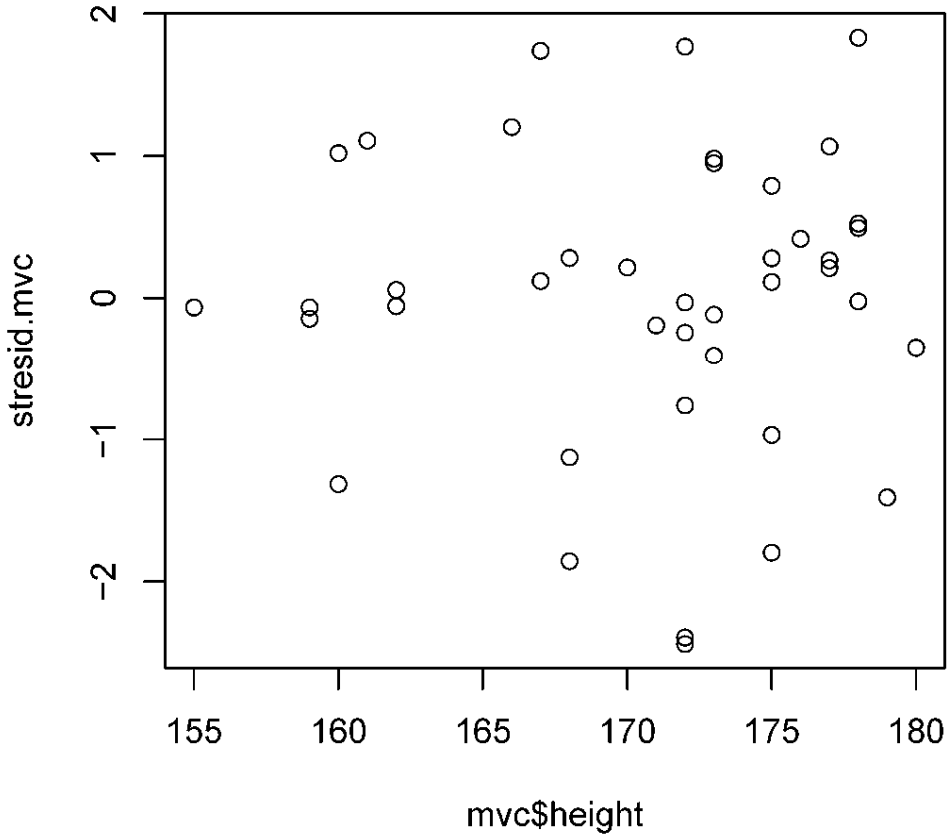
```
> plot(mvc$age, stresid.mvc)
```



- no obvious pattern across age

MVC Residual plots – against predictors (height)

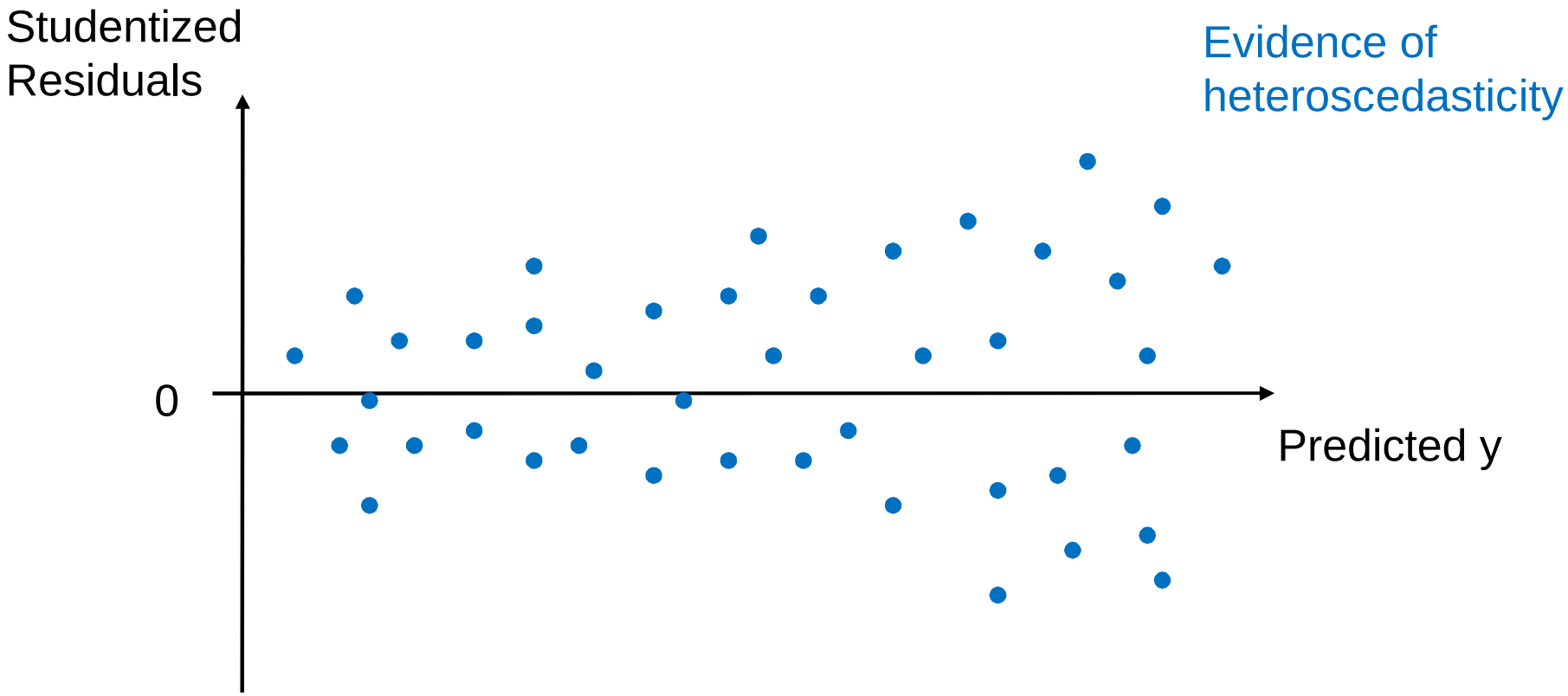
```
> plot(mvc$height, stresid.mvc)
```



- no obvious pattern across height
- note that these are the same residuals, but ordered differently

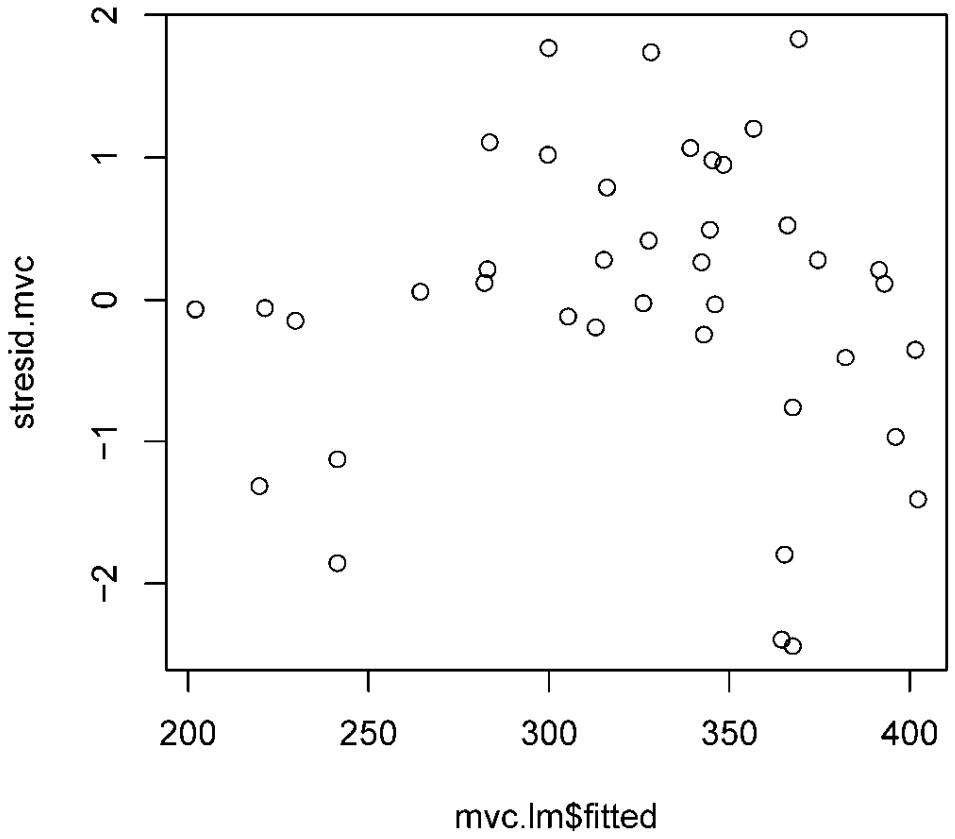
Identify potential heteroscedasticity of the errors

- Plot studentized residuals against fitted values
  - may need to transform the response variable if variance changes with higher fitted values



Residual plot against fitted values

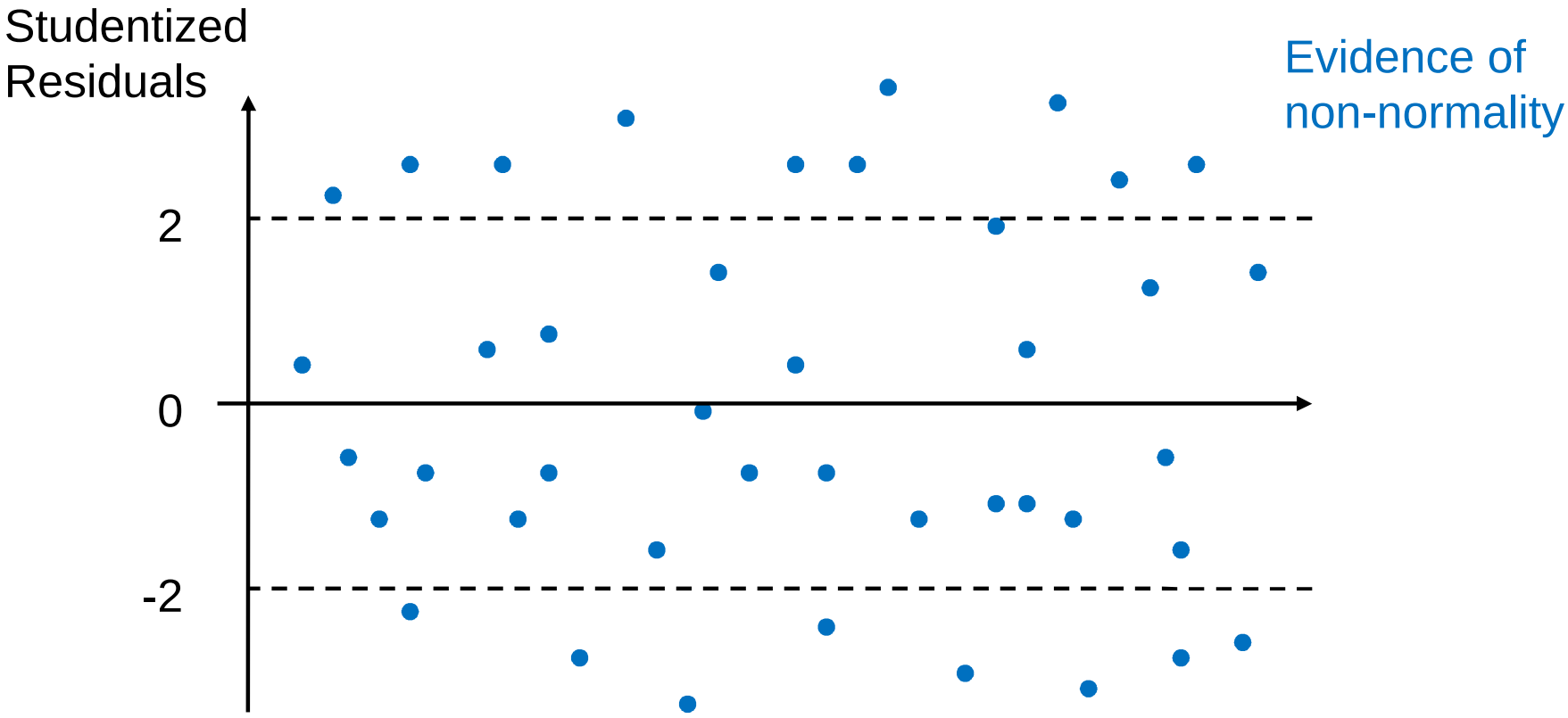
```
> plot(mvc.lm$fitted, stresid.mvc)
```



- no obvious change in variance

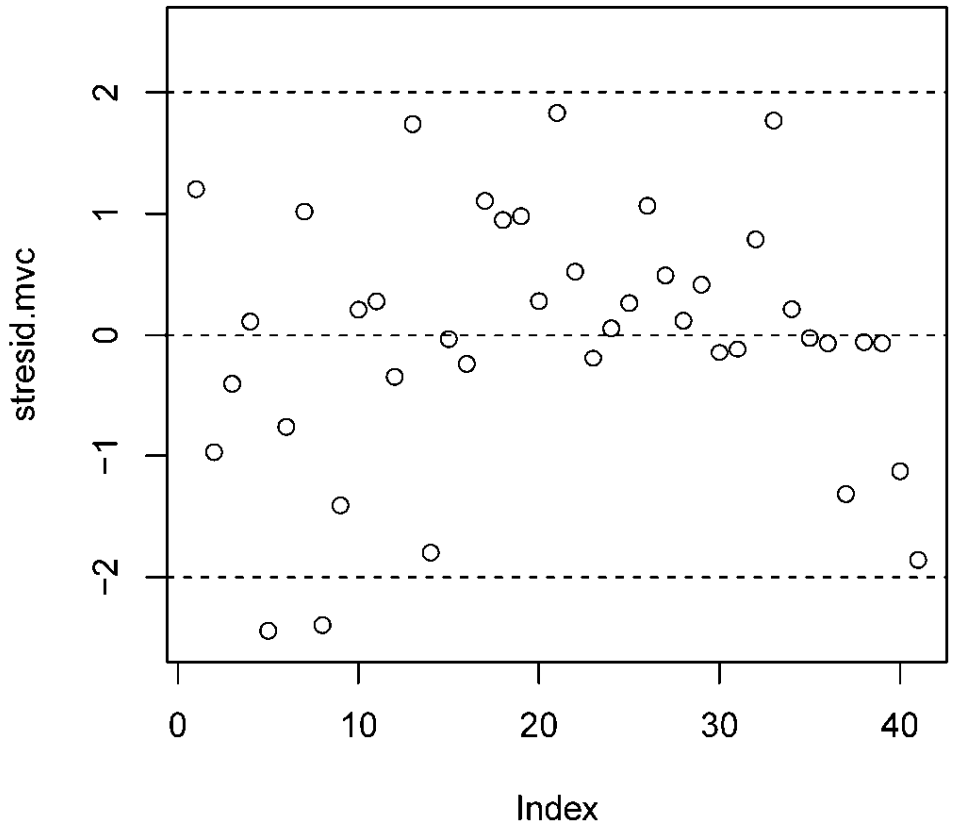
Identify potential non-normality of the errors

- Plot studentized residuals, if errors are normally distributed:
  - Should have a mean zero and variance one
  - Should be symmetrically distributed
  - Should have around 5% of the absolute studentized residuals exceeding 2



MVC Residual plots – studentized residuals

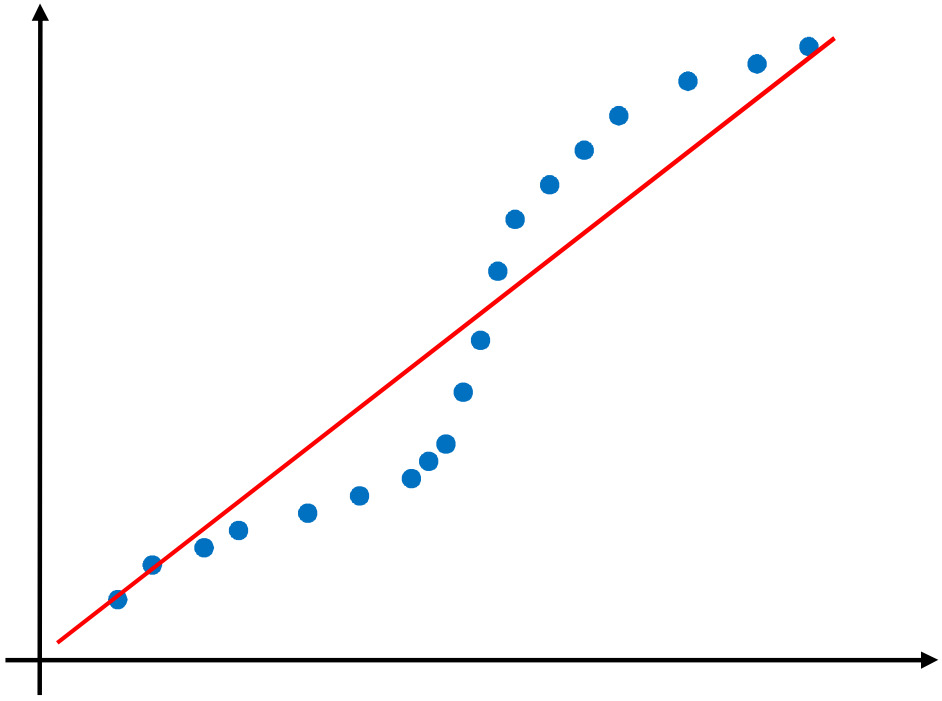
```
> plot(stresid.mvc, ylim=c(-2.5,2.5))
> abline(h=c(0,-2,2), lty=2)
```



- about 5% of the studentized residuals fall outside the range [-2, 2] if the model is true

Identify potential nonnormality of the errors

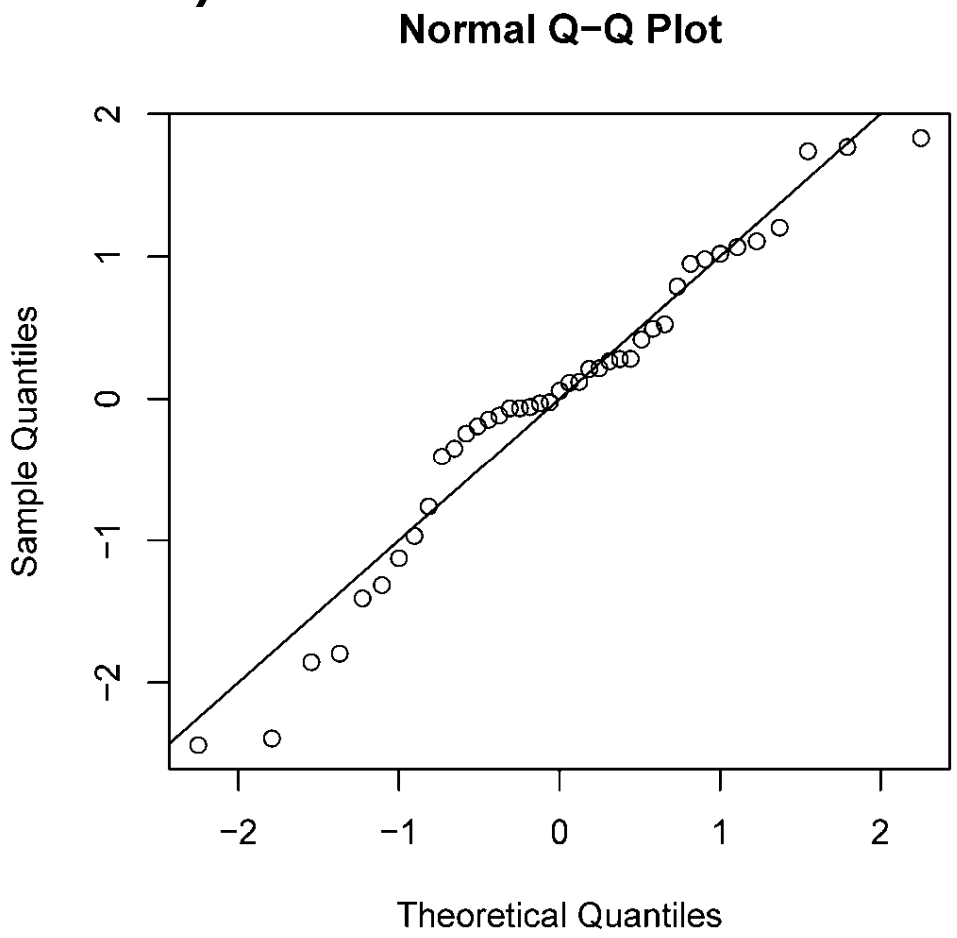
- Normal probability plot (p-p plot) / Normal quantile plot (q-q plot) of residuals
  - If two variables have the same distribution, points should fall on the 45° line on a p-p or q-q plot
  - p-p plot more sensitive to deviation from normality in the middle range; q-q plot more sensitive to outer range



Evidence of non-normality

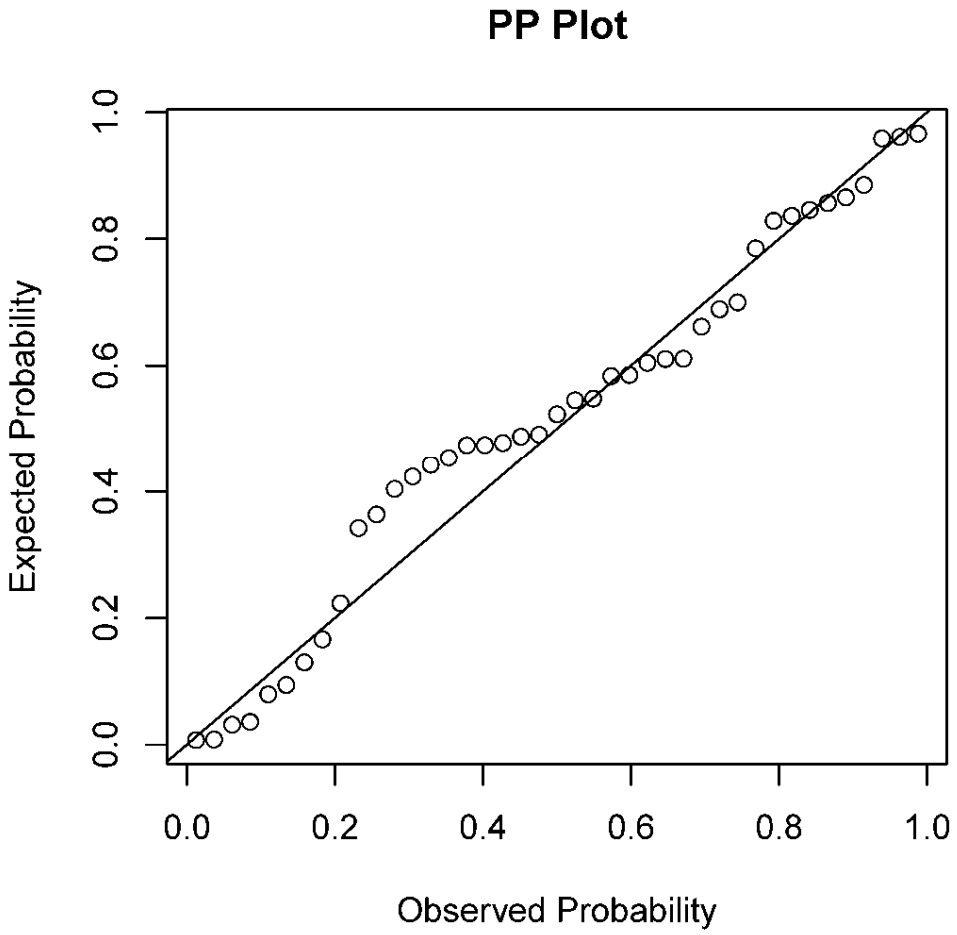
Q-Q Plot for MVC model

```
> qqnorm(stresid.mvc)
> abline(0,1)
```



- no severe deviation from normality

P-P Plot for MVC model



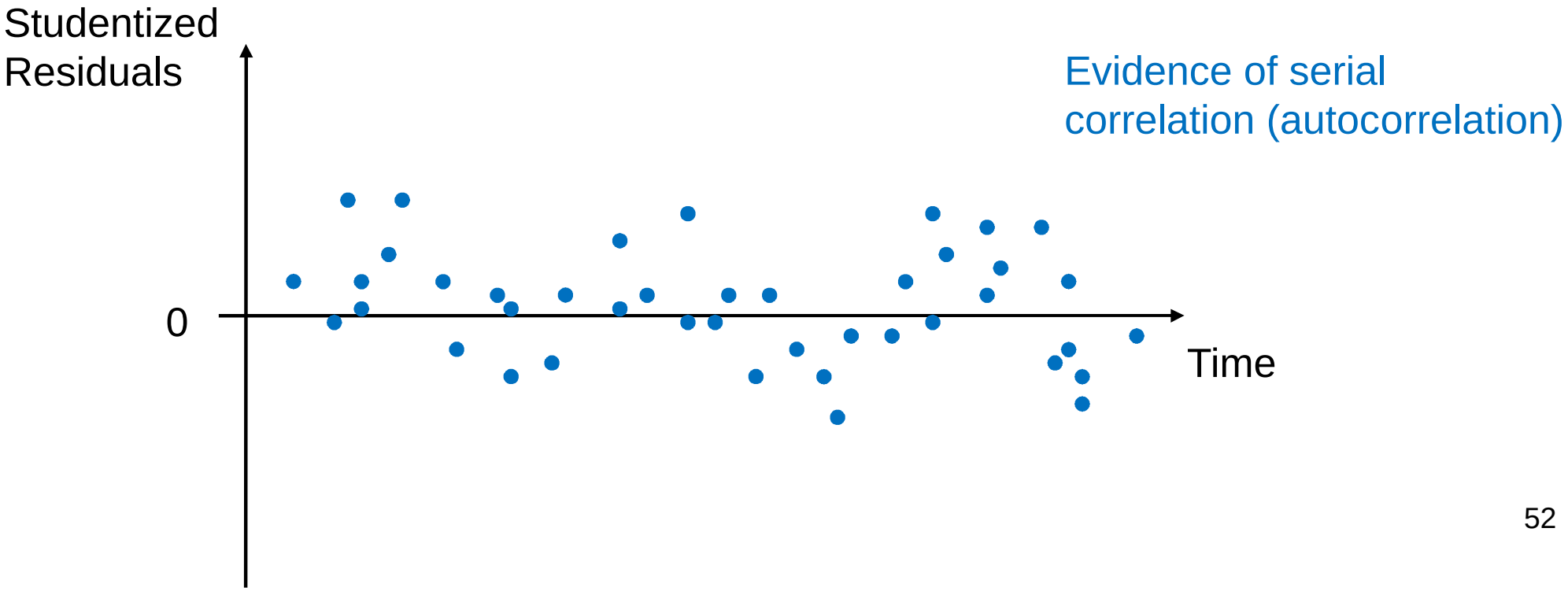
- no severe deviation from normality

(for reference only)

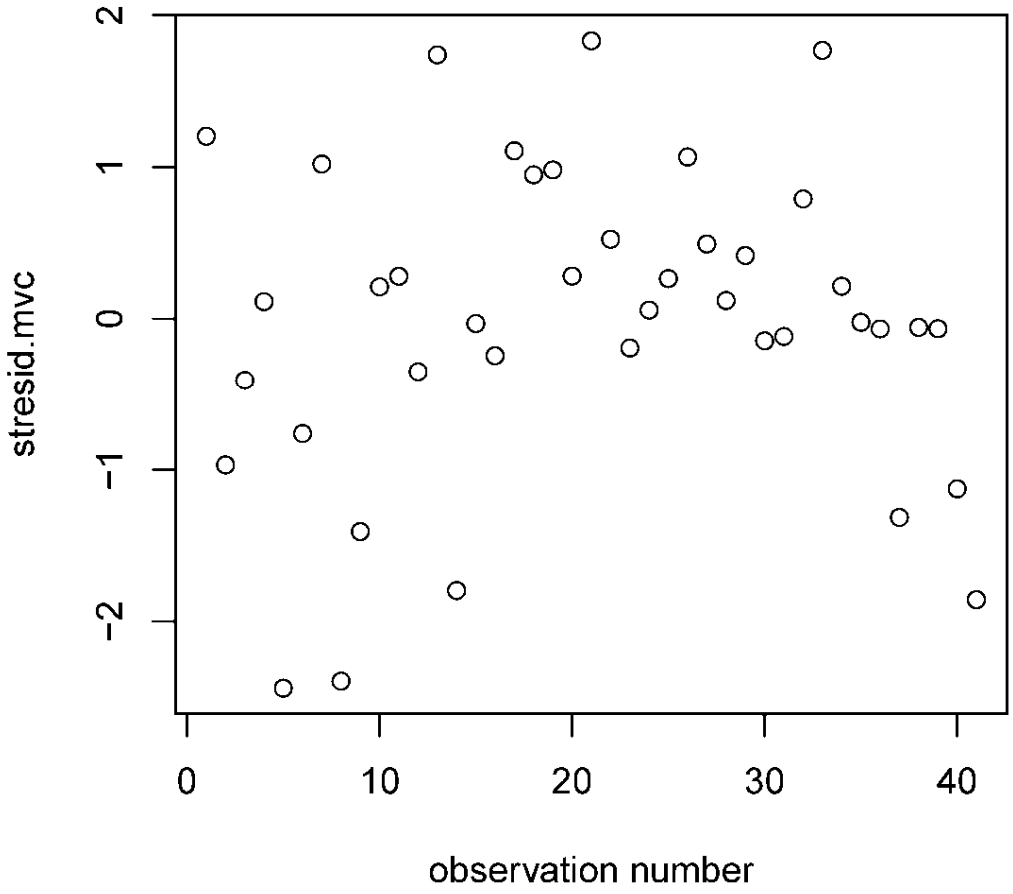
```
> pd.stresid.mvc <-
   pnorm(stresid.mvc)
> plot(ppoints(length(stresid.mvc)),
   sort(pd.stresid.mvc), main = "PP
   Plot", xlab = "Observed
   Probability", ylab = "Expected
   Probability")
> abline(0,1)
```

Identify potential non-independence of the errors

- A common form of non-independence is serial correlation
  - Sometimes a by-product of misspecification of the time trend when time is a predictor in the model
- Plot studentized residuals over time
  - Adjacent errors tend to have the same sign if errors are serially correlated



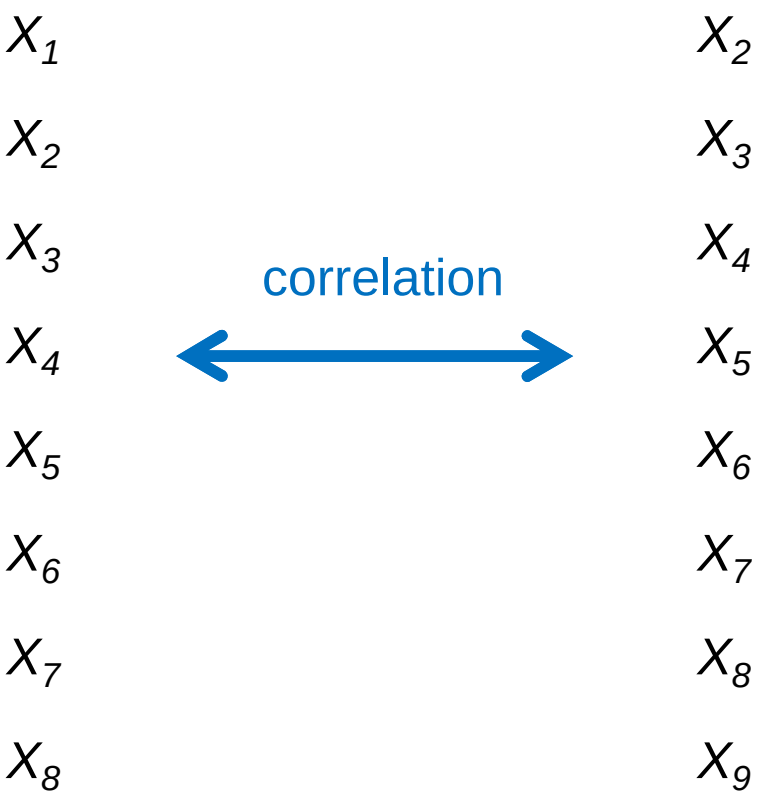
MVC residuals by observation



- Meaningful only if observations were made sequentially
- No obvious autocorrelation in the residuals

Identify serial correlation using ACF plot

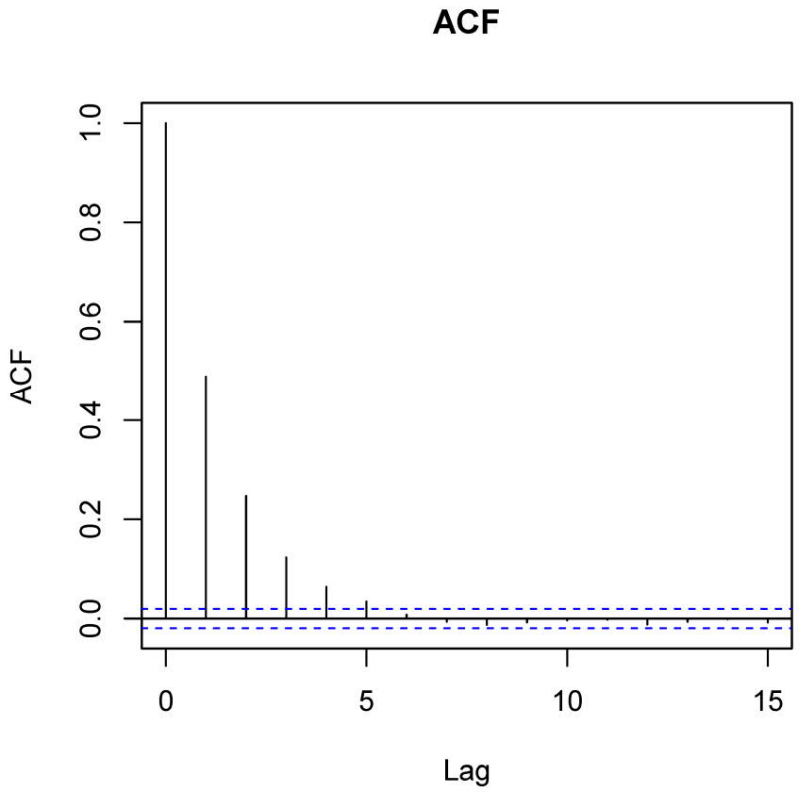
- ACF: autocorrelation function
  - Correlation between the same variable at different time point
  - Define X_t: variable X observed at time t

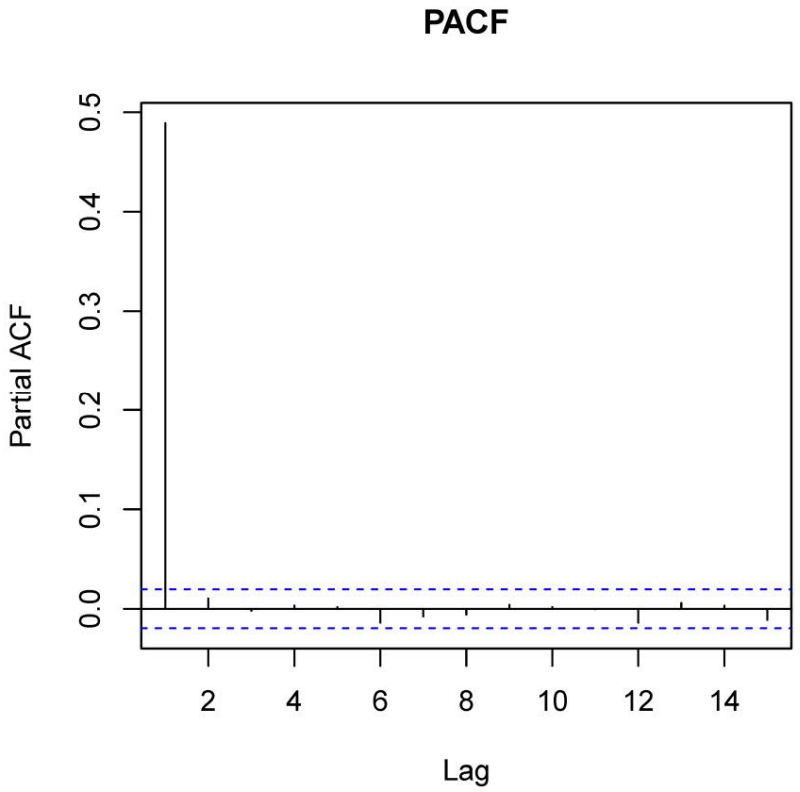


Identify serial correlation using ACF plot

- ACF / partial ACF exceeding confidence limit in the first few lags indicate serial correlations
- Function in R: acf / pacf

Try to produce ACF, PACF plots for the residuals from MVC regression



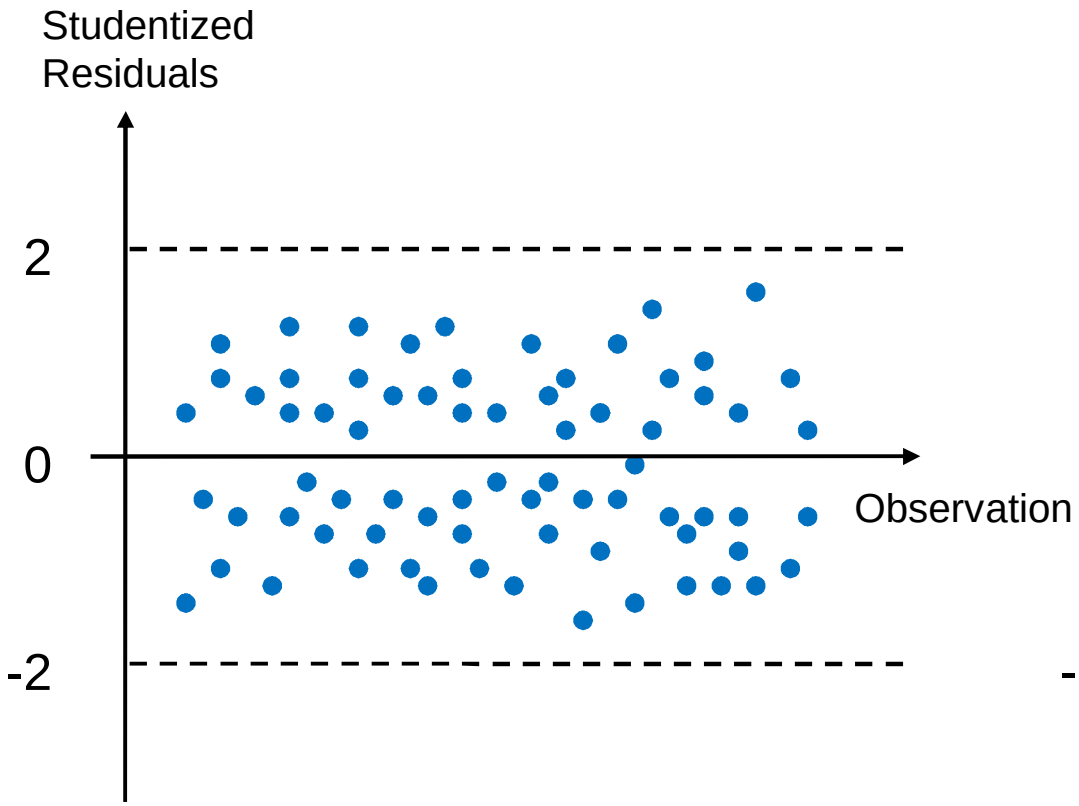


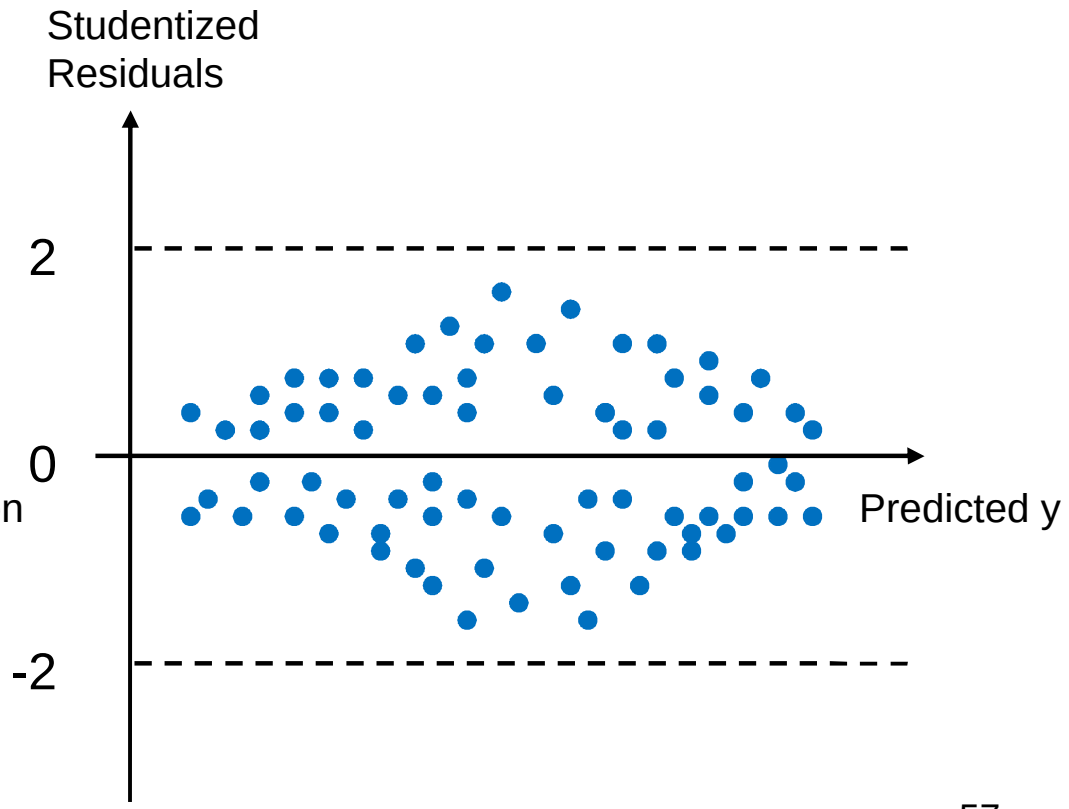
Consequences of assumption violations for linear regression

Problem	Biased coefficients	Biased standard error	Invalid t-test for coefficients
Nonlinear relationship	✓	✓	✓
Heteroscedasticity of the errors		✓	✓
Autocorrelated errors		✓	✓
Non-normality of the errors			✓

In-class exercise: residual plots

- Suppose a linear regression model is fitted
- Do you identify any potential problem from the following residual plots?





Key assumptions for GLM

- Linearity
  - Linear relationship between predictors and dependent variable through a smooth invertible link function
- Errors follow a distribution from the exponential family (e.g. Normal, Bernoulli, Poisson)
- Independence
  - No correlation / autocorrelation between the errors

Model diagnostics for GLM

- Crude residuals are not used, as the variance depends on mean
- Deviance residuals can be used
  - A measure of deviation from the likelihood of a saturated model (no. of parameters = no. of data points)
- Can be obtained by *glm.model$res*
  - (The default are deviance residuals for glm model)
- Studentized deviance residuals can be obtained by rstudent(*glm.model*)
  - follow a standard normal distribution if the model is correctly specified

Multicollinearity

- Collinearity refers to strong linear dependency (high correlation) between two predictor variables
  - e.g. personal income and household income
  - Collinear variables give similar information
- For perfect collinearity between two predictor variables X_1 and X_2
  - We can find constants a, b so that

$$X_1 = a + bX_2$$

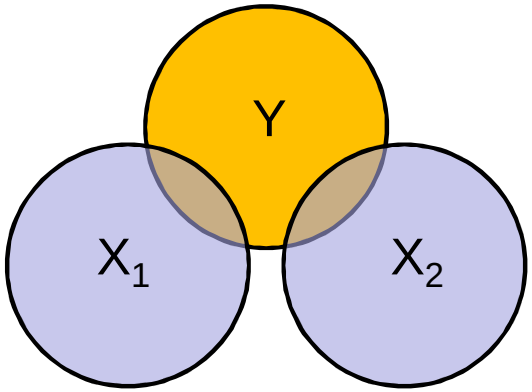
  - e.g. X_1 = degree in Celsius, X_2 = degree in Fahrenheit
- Usually correlation > 0.8 indicates severe collinearity problem
- Multicollinearity:
  - Predictor variable strongly depends on other predictor variables linearly
  - e.g. $X_4 = a + b_1X_1 + b_2X_2 + b_3X_3$ has a very good fit

Problems with multicollinearity

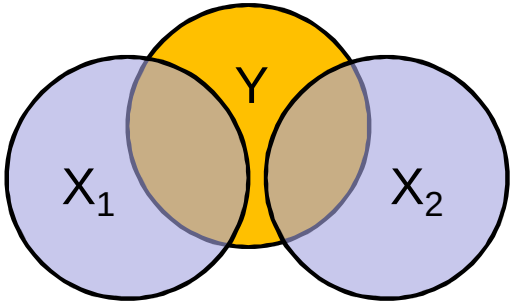
- Multicollinearity does not violate any of the assumptions of linear regression model
- But inflates standard error of the estimated coefficients
  - e.g. for perfectly correlated predictors $X_1 = a + bX_2$
  - $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 = (\alpha + a\beta_1) + (b\beta_1 + \beta_2)X_2$
  - We can always find different sets of (α, β_1, β_2) to represent exactly the same relation
  - For highly correlated predictors, it will be difficult to distinguish their effect on the outcome variable

  → the uncertainty in those estimated coefficients will be large

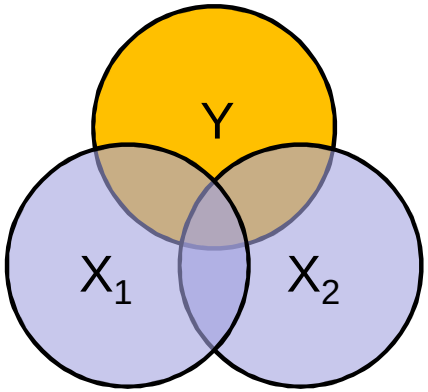
Venn diagram illustration of multicollinearity



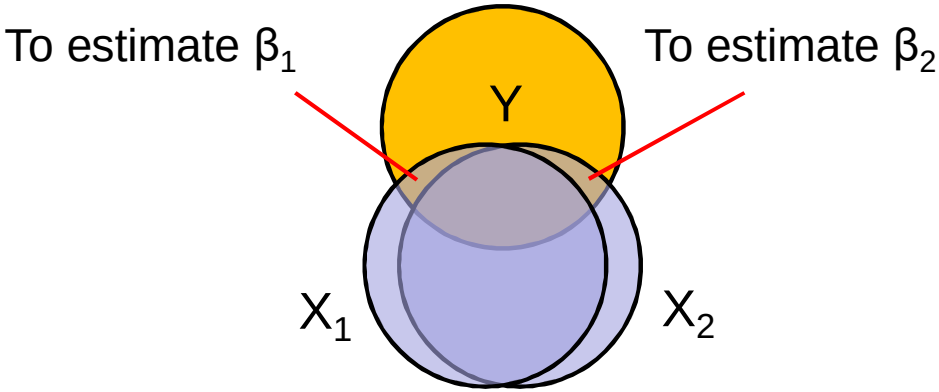
Independent predictors, weak fit



Independent predictors, strong fit



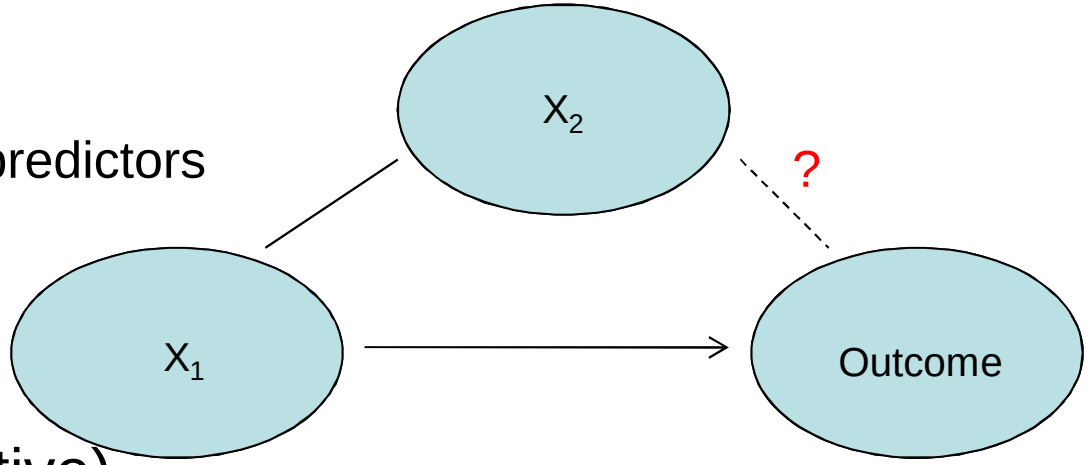
Mildly collinear predictors, moderate fit



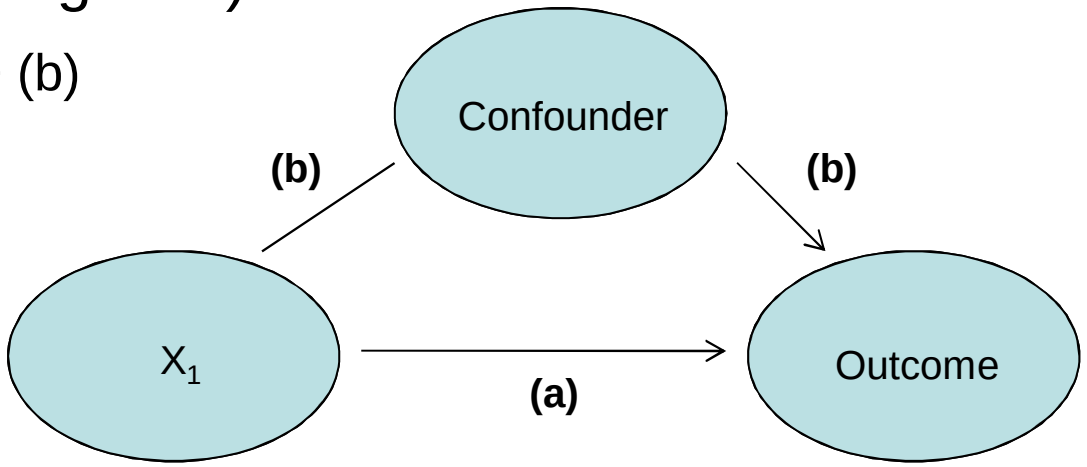
Strongly collinear predictors, moderate fit, larger uncertainty in the effect of X_1 or X_2

Multicollinearity versus confounding

- Uncommon for all predictors to be approximately independent
  - Especially when confounders are present
- Multicollinearity
  - Concern with correlation among predictors



- Confounding (positive or negative)
  - Observed association: (a) + (b)
  - To estimate: (a)
  - Exclusion of important confounders leads to biased estimation of (a)



Multicollinearity diagnostics

- Scatterplot between all predictor variables
  - More useful to identify pairwise collinearity
  - Less helpful to identify multicollinearity of multiple predictors
- Variance inflation factor (VIF)
  - $VIF_j = 1 / (1 - R_j^2)$
  - R_j^2 is the R^2 when predictor X_j is regressed on all other predictors
  - A value of VIF > 10 indicates potential multicollinearity problem

Dealing with multicollinearity

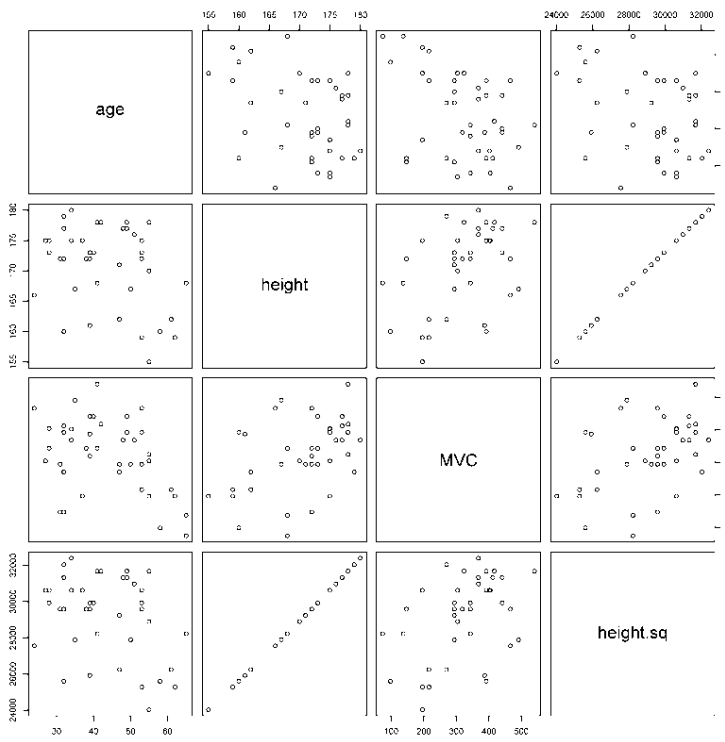
- Do nothing
  - Estimated coefficients are still unbiased
  - but inefficient estimation
- Increase the sample size
- Polynomial terms: centering variables about their mean
  - By subtracting the mean values from the variable (useful for polynomial and interactions only)
  - i.e. $x \rightarrow x - \bar{x}$ $\quad$ $x^2 \rightarrow (x - \bar{x})^2$
- Drop one or more variables causing multicollinearity
  - Most appropriate when two variables basically measure the same thing
  - Problematic if the variable is an important confounder

MVC example

- Suppose we create a new variable ‘height.sq’ which is the square of the heights
- We fit the linear regression model including height.sq:

```
> mvc$height.sq <- mvc$height^2
> pairs(mvc)
```

- height and height.sq highly correlated



MVC regression with highly correlated variables

```
> summary(lm(MVC~age+height+height.sq, data=mvc))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4391.6093 10766.3438   0.408   0.6857
age            -3.1422     1.4903  -2.108   0.0418 *
height        -52.3026   127.8036  -0.409   0.6847
height.sq       0.1712     0.3791   0.452   0.6542
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 99.97 on 37 degrees of freedom
Multiple R-squared:  0.2653,    Adjusted R-squared:  0.2057
F-statistic: 4.453 on 3 and 37 DF,  p-value: 0.009074
```

VIF in R

- Install and load the “car” package

```
> require(car)
> mvc.lm3 <- lm(MVC~age+height+height.sq, data=mvc)
> vif(mvc.lm3)
       age      height   height.sq
  1.140494 2788.043855 2784.139173
```

- Very large VIF for the height terms
- For polynomial and interactions, VIF can be reduced by centering
- For other collinearity problem, centering does not help.

Centering in R

```
> mvc$ct.height <- scale(mvc$height, scale=F)
> mvc$ct.height.sq <- mvc$ct.height^2
> mvc.lm4 <- lm(MVC~age+ct.height+ct.height.sq, data=mvc)
> vif(mvc.lm4)
       age    ct.height ct.height.sq
  1.140494     1.603104     1.534462
```

Before centering:

```
> vif(mvc.lm3)
       age      height   height.sq
  1.140494 2788.043855 2784.139173
```

MVC example: regression after centering

```
> summary(mvc.lm4)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  451.7998    67.0772   6.736  6.4e-08 ***
age           -3.1422     1.4903  -2.108   0.0418 *
ct.height      6.1506     3.0646   2.007   0.0521 .
ct.height.sq   0.1712     0.3791   0.452   0.6542
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 99.97 on 37 degrees of freedom
Multiple R-squared:  0.2653,    Adjusted R-squared:  0.2057
F-statistic: 4.453 on 3 and 37 DF,  p-value: 0.009074
```

- Note that R^2 is the same as before

Other potential problems in regression models

- Multilevel structure
  - e.g. correlation within hospital or ward
  - Multi-level models
- Measurement error
  - Imprecise measurement in predictors will attenuate estimated coefficients toward zero

Review

- We have seen how to describe, present and analyse count data with Poisson and negative binomial regression
- We have discussed the way to handle overdispersed data
- We have seen how to interpret and present the results of count regression models in terms of absolute and relative risks
- We have discussed how to identify and handle influential observations
- We have discussed model diagnostics
- We have discussed how to identify and handle multicollinearity problem

Further reading

- Gelman, A., Hill, J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2007
- Long, J. S. *Regression models for categorical and limited dependent variables*. Sage Publications, 1997
- Vittinghoff, E, et al. *Regression methods in biostatistics.* Springer, 2005