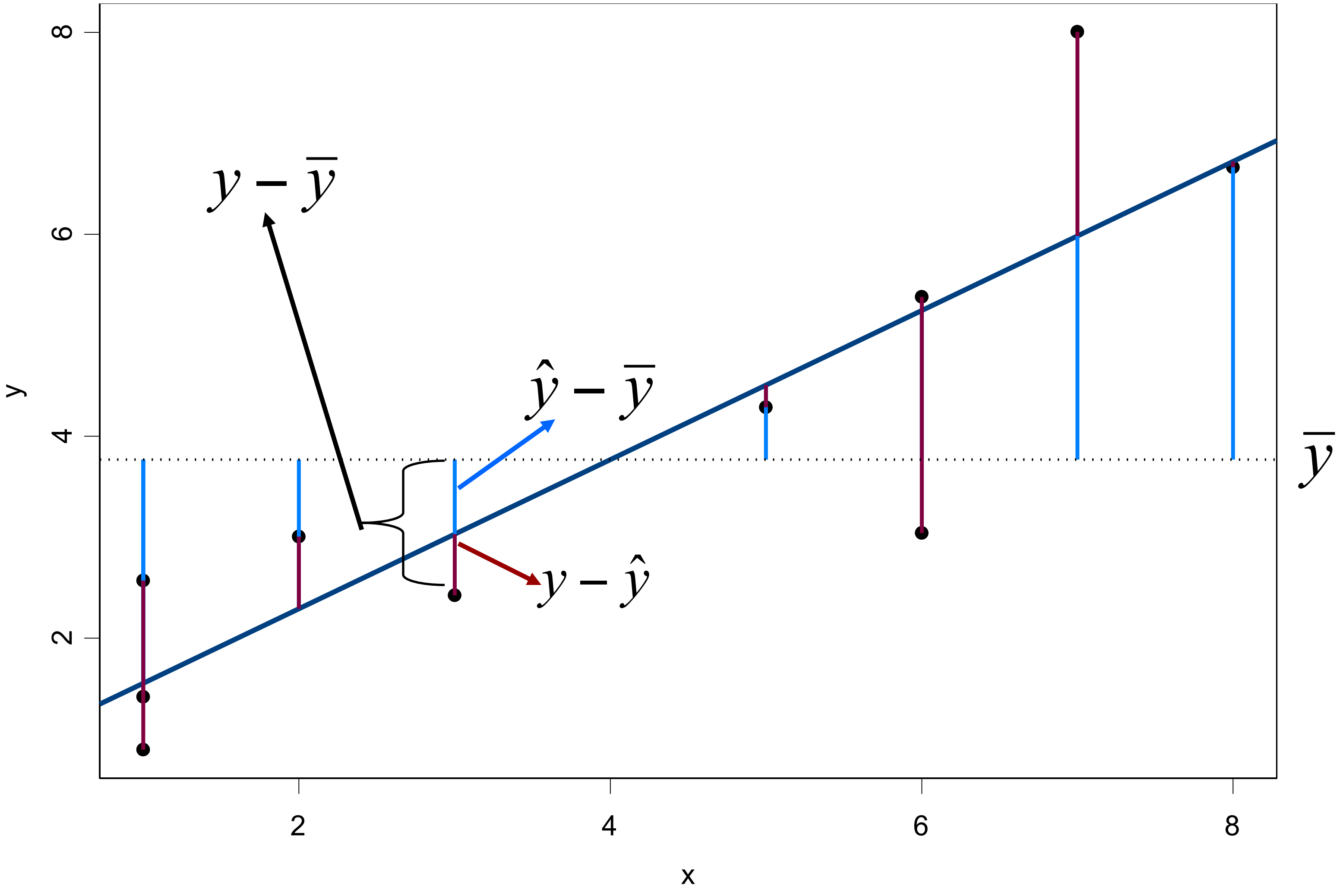
Simple Linear Regression: R^2

- Given no linear association:
  - We could simply use the sample mean to predict E(Y). The variability using this simple prediction is given by SST (to be defined shortly).

- Given a linear association:
  - The use of *X* permits a potentially better prediction of Y by using E(Y|X).
  - **Question:** What did we gain by using *X*?

Let's examine this question with the following figure

Decomposition of sum of squares



Decomposition of sum of squares

It is always true that: $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$

It can be shown that:

$$\boxed{\begin{array}{c} \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\ \\ SST \quad = \quad SSE \quad + \quad SSR \end{array}}$$

SST: describes the total variation of the Y_i.

SSE: describes the variation of the Y_i around the regression line.

SSR: describes the structural variation; how much of the variation is due to the regression relationship.

This decomposition allows a characterization of the usefulness of the covariate X in predicting the response variable Y.

Simple Linear Regression: R^2

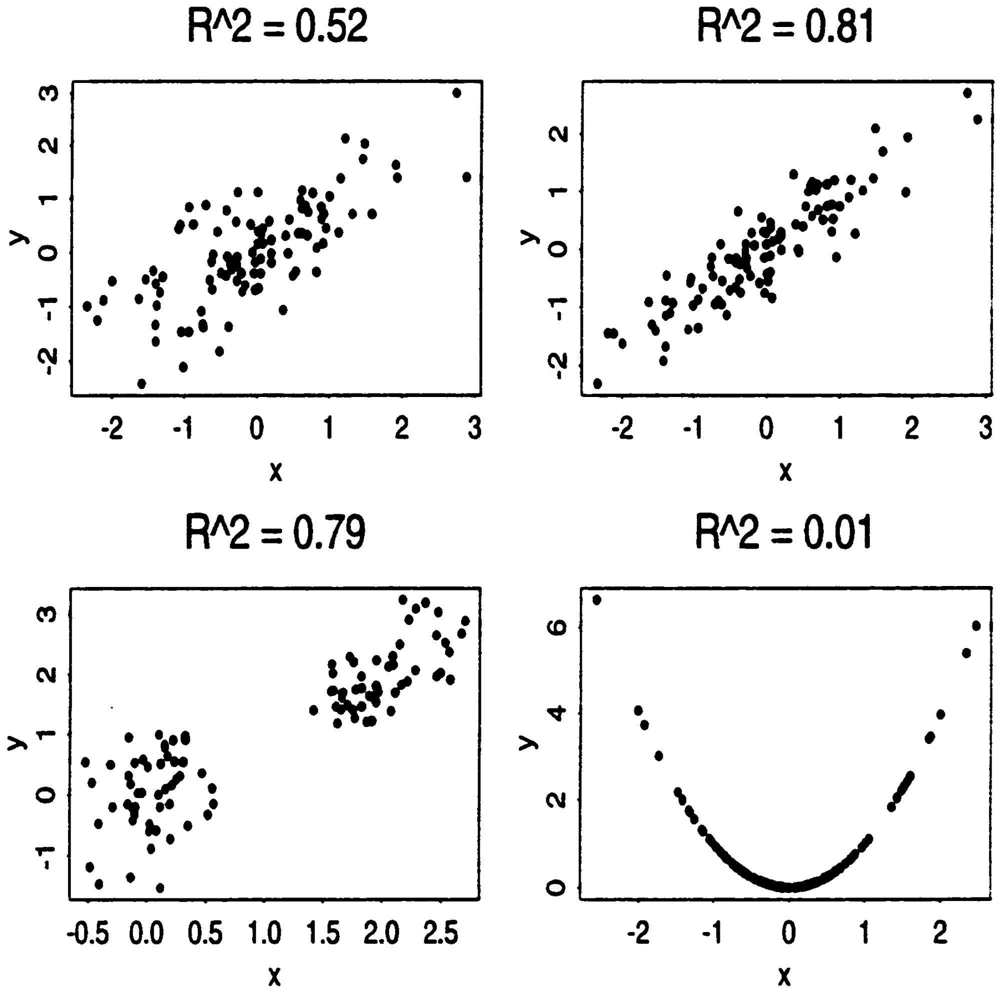
- Given no linear association:
  - We could simply use the sample mean to predict E(Y). The variability between the data and this simple prediction is given as SST.

- Given a linear association:
  - The use of X permits a potentially better prediction of Y by using E(Y| X).
  - **Question:** What did we gain by using X?
  - **Answer:** We can answer this by computing the proportion of the total variation that can be explained by the regression on X

$$\boxed{R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}}$$

- This R^2 is, in fact, the correlation coefficient squared.

Examples of R^2



Low values of R^2 indicate that the model is not adequate. However, high values of R^2 do not mean that the model is adequate!!

Cholesterol Example:

Scientific Question: Can we predict cholesterol based on age?

```
> fit = lm(chol ~ age)
> summary(fit)

Call:
lm(formula = chol ~ age)

Residuals:
      Min        1Q    Median        3Q       Max
-60.45306 -14.64250  -0.02191  14.65925  58.99527

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 166.90168    4.26488  39.134  < 2e-16 ***
age           0.31033    0.07524   4.125 4.52e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.69 on 398 degrees of freedom
Multiple R-squared: 0.04099,      Adjusted R-squared: 0.03858
F-statistic: 17.01 on 1 and 398 DF,  p-value: 4.522e-05
```

```
> confint(fit)
                  2.5 %      97.5 %
(Intercept) 158.5171656 175.2861949
age           0.1624211   0.4582481
```

Cholesterol Example:

Scientific Question: Can we predict cholesterol based on age?

- $R^2=0.04$
- What does R^2 tell us about our model for cholesterol?

Cholesterol Example:

Scientific Question: Can we predict cholesterol based on age?

- $R^2=0.04$
- What does R^2 tell us about our model for cholesterol?
- Answer: 4% of the variability in cholesterol is explained by age. Although mean cholesterol increases with age, there is much more variability in cholesterol than age alone can explain

Cholesterol Example:

Scientific Question: Can we predict cholesterol based on age?

- Decomposition of Sum of Squares and the F-statistic

Degrees of freedom

Decomposition of the Sum of Squares

Mean Squares: SS/df

F-statistic: MSR/MSE

```
> anova(fit)
Analysis of Variance Table

Response: chol
              Df Sum Sq Mean Sq F value    Pr(>F)
SSR=age        1   8002  8001.7  17.013 4.522e-05 ***
SSE=Residuals 398 187187   470.3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In simple linear regression:

F-statistic = (t-statistic for slope)^2

Hypothesis being tested: H_0: $\beta_1=0$, H_1: $\beta_1\neq0$.

Simple Linear Regression: Assumptions

1. E[Y|x] is related linearly to x
2. Y’s are independent of each other
3. Distribution of [Y|x] is normal
4. Var[Y|x] does not depend on x

Linearity
Independence
Normality
Equal variance

Can we assess if these assumptions are valid?

Model Checking: Residuals

- **(Raw or unstandardized) Residual**: difference (r_i) between the observed response and the predicted response, that is,

$$r_i = y_i - \hat{y}_i$$
$$= y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

The residual captures the component of the measurement y_i that cannot be "explained" by x_i.

Model Checking: Residuals

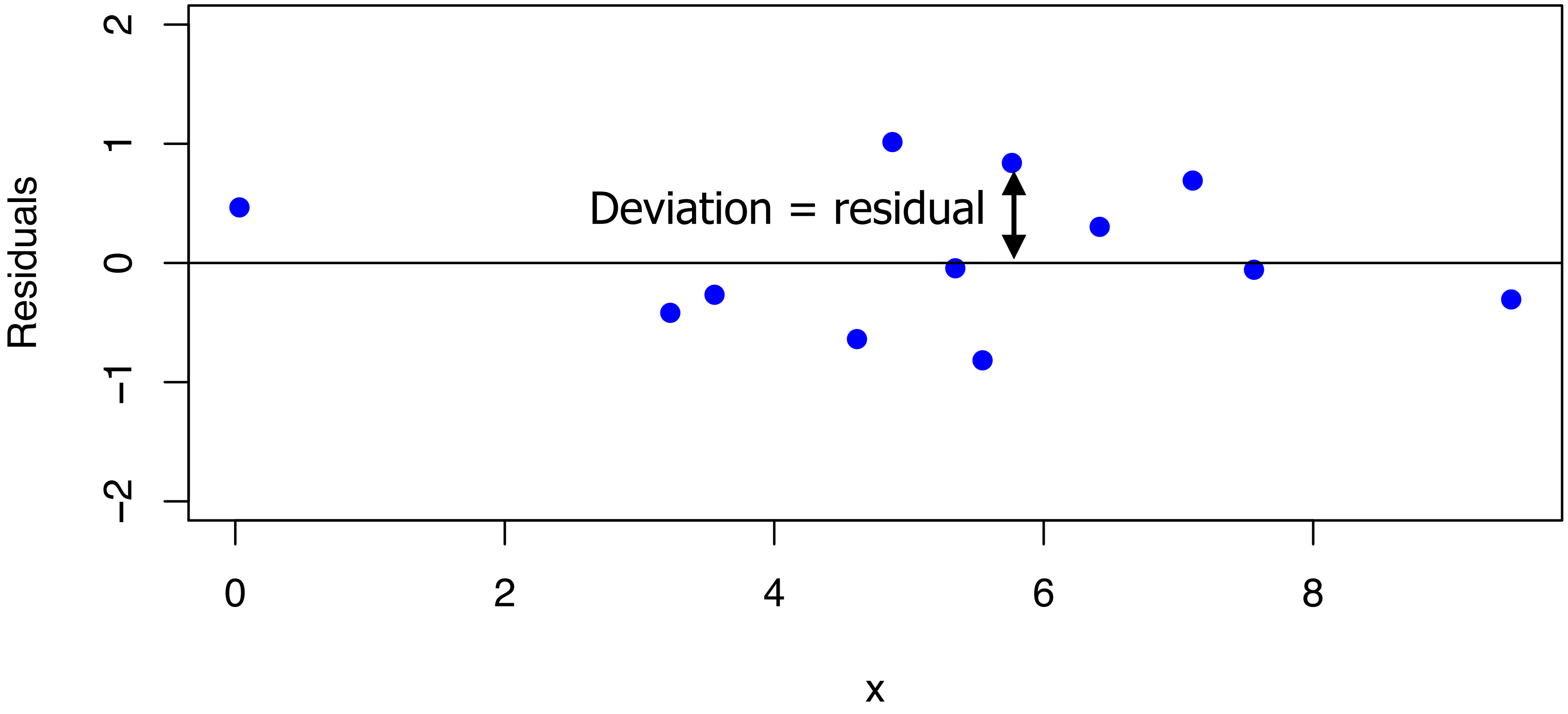
- Residuals can be used to
  - Identify poorly fit data points
  - Identify unequal variance (heteroscedasticity)
  - Identify nonlinear relationships
  - Identify additional variables
  - Examine normality assumption

Model Checking: Residuals

Linearity	Plot residual vs X or vs $\hat{Y}$ Q: Is there any structure?
Independence	 Q: Any scientific concerns?
Normality	Residual histogram or qq-plot Q: Symmetric? Normal?
Equal variance	Plot residual vs X Q: Is there any structure?

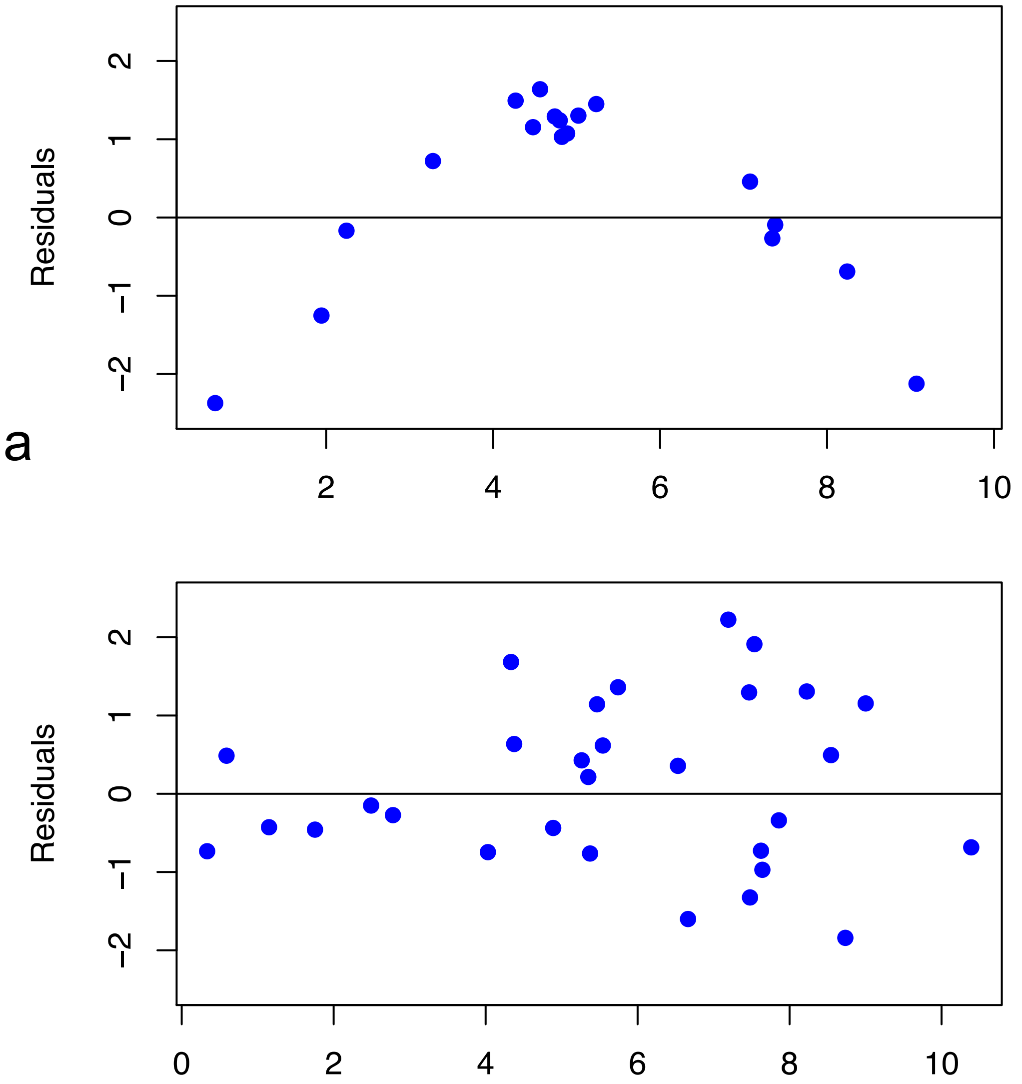
Model Checking: Residuals

- If the linear model is appropriate we should see an **unstructured horizontal band of points centered at zero** as seen in the figure below



Model Checking: Residuals

The model does not provide a good fit in these cases!



Violations of the model assumptions? How?

Linearity

- The linearity assumption is important: interpretation of the slope estimate depends on the assumption of the same rate of change in E(Y|X) over the range of X
- Preliminary Y-X scatter plots and residual plots can help identify non-linearity
- If linearity cannot be assumed, consider alternatives such as polynomials, fractional polynomials, splines or categorizing X

Independence

- The independence assumption is also important: whether observations are independent will be known from the study design

- There are statistical approaches to accommodate dependence, e.g. dependence that arises from cluster designs

Normality

- The Normality assumption can be visually assessed by a histogram of the residuals or a normal QQ-plot of the residuals

- A QQ-plot is a graphical technique that allows us to assess whether a data set follows a given distribution (such as the Normal distribution)
  - The data are plotted against a given theoretical distribution
    - Points should approximately fall in a straight line
    - Departures from the straight line indicate departures from the specified distribution.

- However, for moderate to large samples, the Normality assumption can be relaxed

See, e.g., Lumley T et al. The importance of the normality assumption in large public health data sets. Annu Rev Public Health 2002; 23: 151-169.

Equal variance

- Sometimes variance of Y is not constant across the range of X (heteroscedasticity)
- Little effect on point estimates but variance estimates may be incorrect
- This may affect confidence intervals and p-values
- To account for heteroscedasticity we can
  - Use robust standard errors
  - Transform the data
  - Fit a model that does not assume constant variance (GLM)

Robust standard errors

- Robust standard errors correctly estimate variability of parameter estimates even under non-constant variance
  - These standard errors use empirical estimates of the variance in y at each x value rather than assuming this variance is the same for all x values

- Regression point estimates will be unchanged

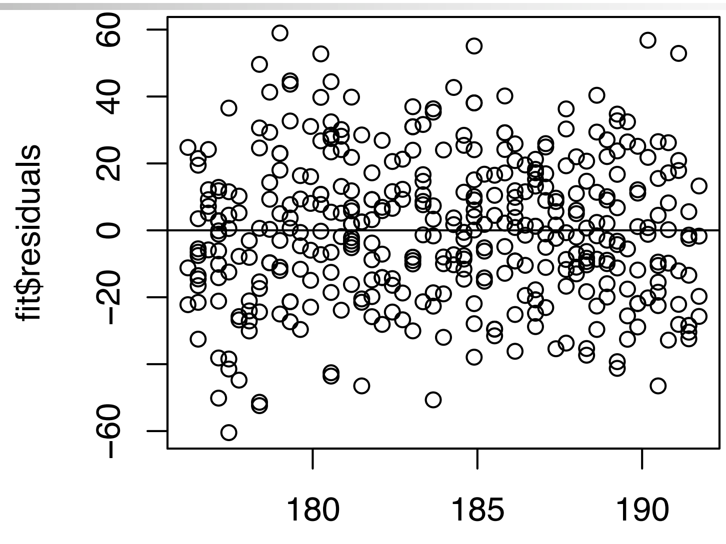
- Robust or empirical standard errors will give correct confidence intervals and p-values

Cholesterol-Age example: Residuals

Plot of residuals versus fitted values

Structure?

Heteroscedasticity?

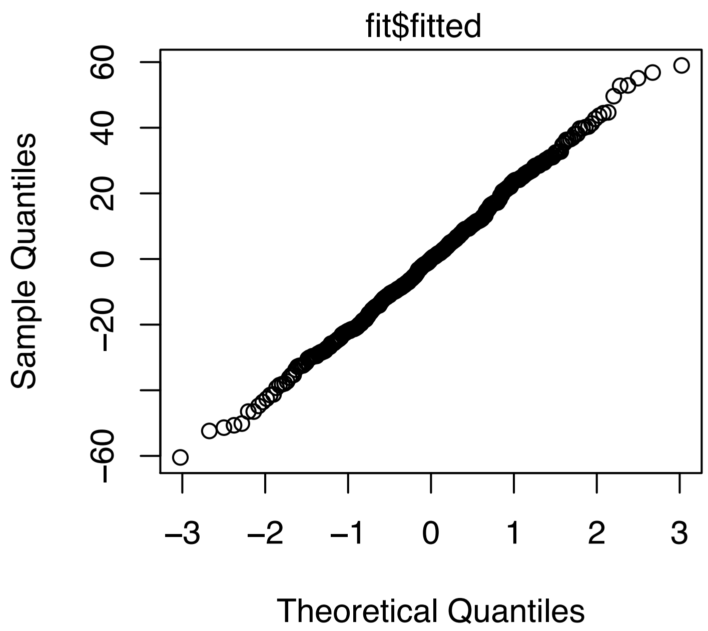


R COMMAND:

plot(fit$fitted, fit$residuals)

Plot of residuals versus quantiles of a normal distribution(for n > 30)

Normality?

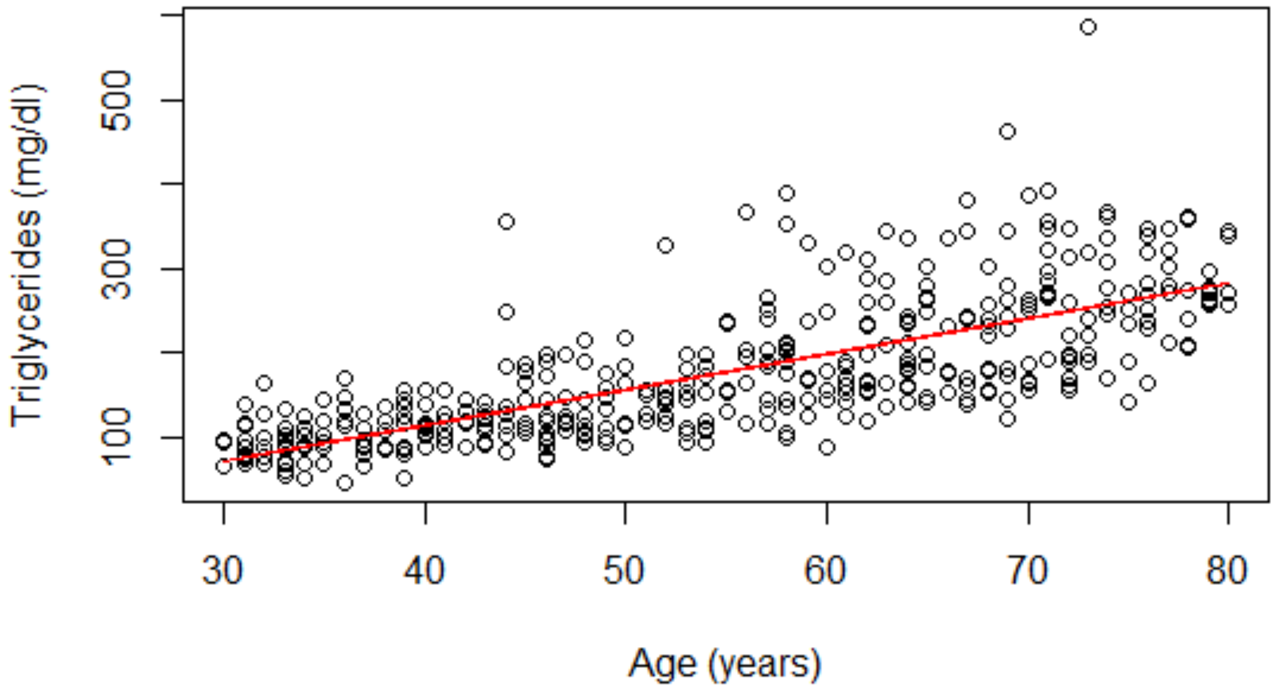


R COMMAND:

qqnorm(fit$residuals)

Another example

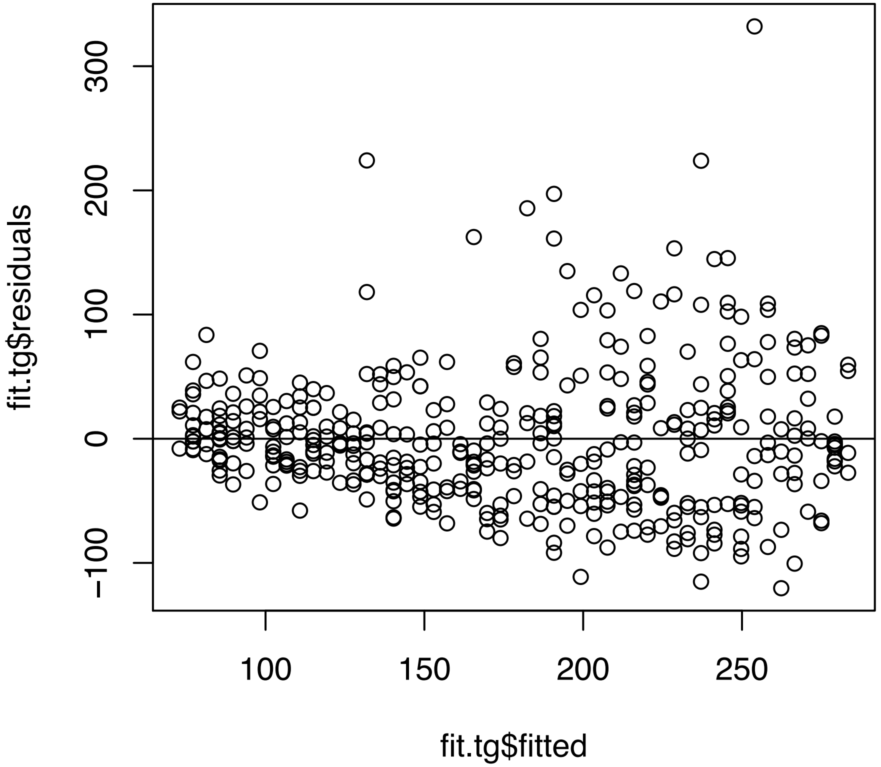
- Linear regression for association between age and triglycerides




```
> fit.tg=lm(TG~age)
```

Robust standard errors

- Residual analysis suggests mean-variance relationship
- Use robust standard errors to get correct variance estimates



Cholesterol example: Robust standard errors

- Linear regression results:

```
> summary(fit.tg)

Call:
lm(formula = TG ~ age)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -53.3059    11.1339  -4.788 2.38e-06 ***
age           4.2090     0.1964  21.429  < 2e-16 ***
```

Point estimates are unchanged

- Results incorporating robust SEs:

```
> fit.tg.robust = coeftest(fit.tg, vcov = sandwich)
> fit.tg.robust

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -53.30593    8.73874  -6.100 2.515e-09 ***
age           4.20896    0.18134  23.211 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cholesterol example: Robust standard errors

- Linear regression results:

```
> summary(fit.tg)

Call:
lm(formula = TG ~ age)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -53.3059    11.1339  -4.788 2.38e-06 ***
age           4.2090     0.1964  21.429  < 2e-16 ***
```

Standard errors are corrected

- Results incorporating robust SEs:

```
> fit.tg.robust = coeftest(fit.tg, vcov = sandwich)
> fit.tg.robust

t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)
(Intercept) -53.30593    8.73874  -6.100 2.515e-09 ***
age           4.20896    0.18134  23.211 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Transformations

- Some reasons for using data transformations
  - Content area knowledge suggests nonlinearity
  - Original data suggest nonlinearity
  - Equal variance assumption violated
  - Normality assumption violated

- Transformations may be applied to the response, predictor or both
  - Be careful with the interpretation of the results

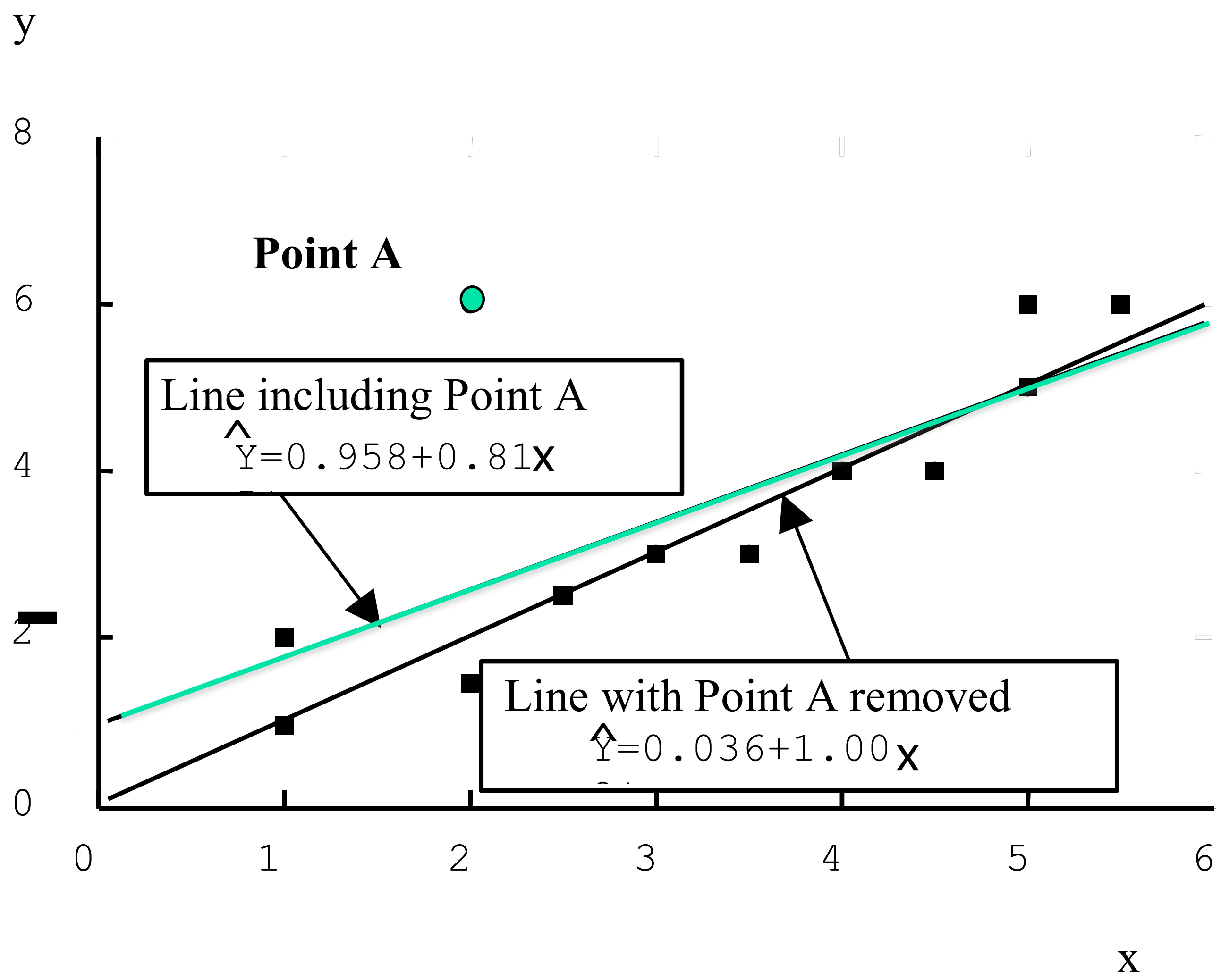
- Rarely do we know which transformation of the predictor provides best "linear" fit – best to choose transformation on scientific grounds
  - As always, there is a danger in using the data to estimate the best transformation to use
    - If there is no association of any kind between the response and the predictor, a "linear" fit (with a zero slope) is the correct one
    - Trying to detect a transformation is thus an informal test for an association
      - Multiple testing procedures inflate the Type I error

Model Checking: Outliers vs Influential observations

- **Outlier:** an observation with a residual that is unusually large (positive or negative) as compared to the other residuals.

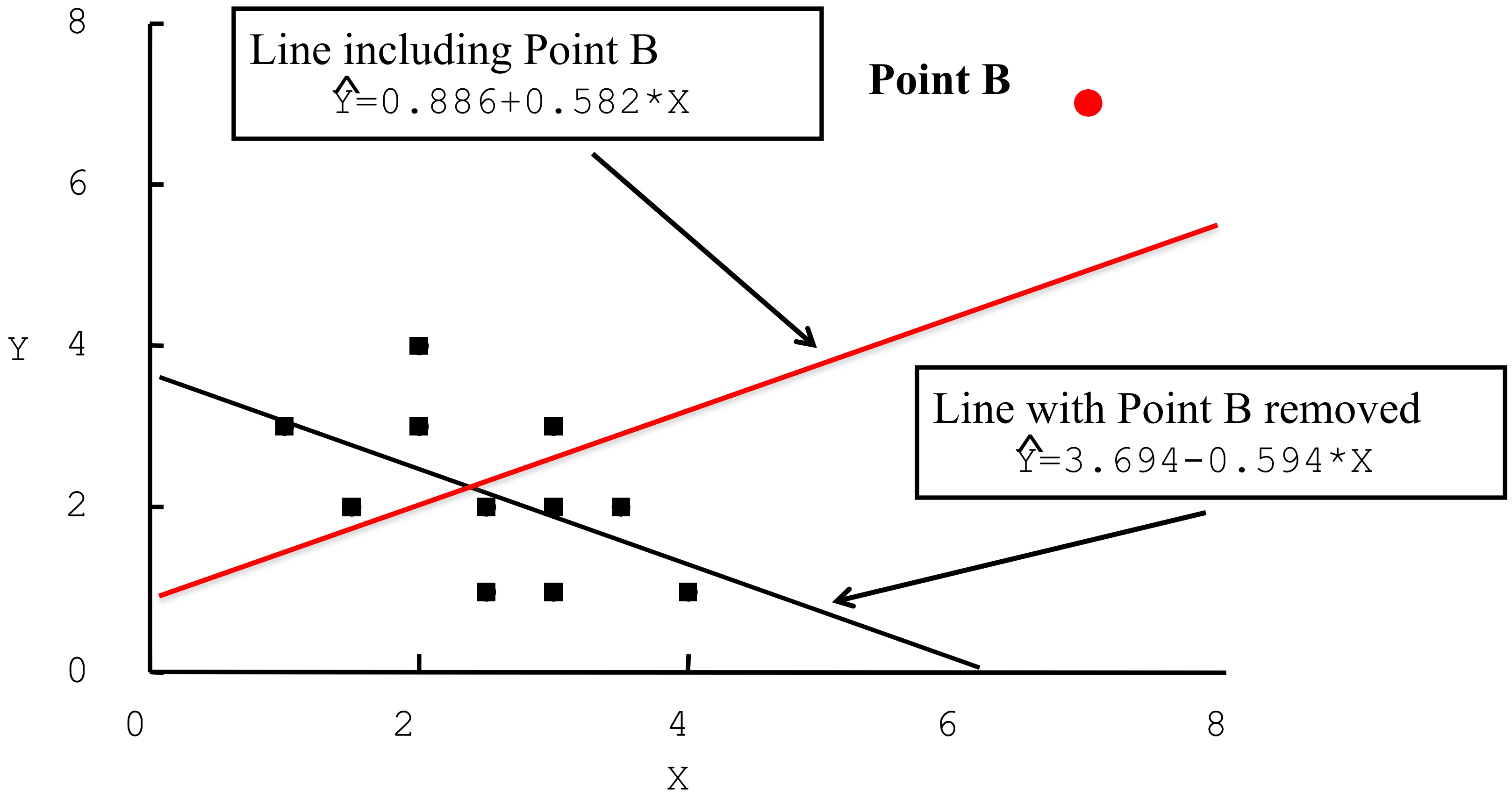
- **Influential point:** an observation that has a notable influence in determining the regression equation.
  - Removing such a point would markedly change the position of the regression line.
  - Observations that are somewhat extreme for the value of x can be influential.

Outlier vs Influential observations



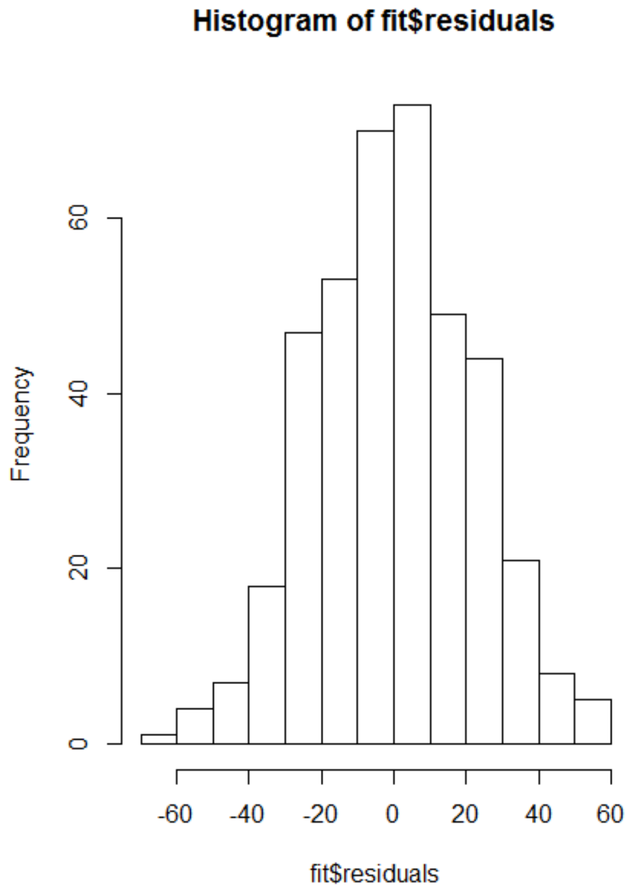
Point A is an *outlier, but is not influential*.

Outlier vs Influential observations



Point B *is influential, but not an outlier.*

Cholesterol-Age Example: Residuals



No extreme outliers

Model Checking: Deletion diagnostics

$$\Delta\beta_{(i)} = \hat{\beta} - \hat{\beta}_{(-i)}$$: Delta-beta

$$\frac{\Delta\beta_{(i)}}{se(\hat{\beta})}$$: Standardized Delta-beta

Delta-beta : tells how much the regression coefficient changed by excluding the i^th observation

Standardized delta-beta : approximates how much the t-statistic for a coefficient changed by excluding the i^th observation

Cholesterol-Age Example: Deletion diagnostics

```
> dfb = dfbeta(fit)
> index=order(abs(dfb[,2]),decreasing=T)
> cbind(dfb[index[1:15],],age[index[1;15]])
 (Intercept)          age
114  -0.9893663  0.015268514 34
166  -0.6827966  0.014888475 78
255  -0.6190643  0.013902713 75
186  -0.8544144  0.013279531 33
113   0.5376293 -0.011943495 76
325  -0.7517511  0.011308451 37
365   0.7676508 -0.011297278 39
257  -0.7374003  0.011092575 37
290  -0.7024787  0.010757541 35
144   0.7120264 -0.010710881 37
197  -0.6784150  0.010469720 34
296  -0.6499386  0.010101515 33
231  -0.6293174  0.009712016 34
7     0.4403297 -0.009524470 79
252  -0.5981020  0.009412761 31
```

No evidence of influential points. The largest (in absolute value)
delta beta is 0.015 compared to the estimate of 0.31 for the regression coefficient.

Model Checking

- What to do if you find an outlier and/or influential observation:

  - Check it for accuracy

  - Decide (based on scientific judgment) whether it is best to keep it or omit it
    - If you think it is representative, and likely would have appeared in a larger sample, keep it
    - If you think it is very unusual and unlikely to occur again in a larger sample, omit it
    - Report its existence [whether or not it is omitted]

Simple Linear Regression: Impact of Violations of Model Assumptions

	Non Linearity	Non Normality	Unequal Variances	Dependence
Estimates	Problematic	Little impact for most departures. Extreme outliers can be a problem.	Little impact	Mostly little impact
Tests/CIs	Problematic	Little impact for most departures. CIs for correlation are sensitive.	Variance estimates may be wrong, but the impact is usually not dramatic	Variance estimates may be wrong
Correction	Choose a nonlinear approach (possible within the linear regression framework)	Mostly no correction needed. Delete outliers (if warranted) or use robust regression	Use robust standard errors	Regression for dependent data

Exercise

- Work on Exercises 4-6
  - Try each exercise on your own
  - Make note of any questions or difficulties you have
  - At **10:30AM PT** we will meet as a group to go over the solutions and discuss your questions