## Residual diagnostics

- Residuals can be plotted against independent/predictor variables to check for model inadequacy. (e.g. if relationship is quadratic, but only a linear model was fit, this plot will reveal a pattern between residuals and predictor.)
- Residuals can be plotted against predicted values to look for inhomogeneity of variance (heteroscedasticity). Look for residuals for which variability increases or "fans out" as one looks left-to-right in this plot (or vice-versa).
- The sorted residuals can be plotted against the normal inverse of the empirical CDF of the residuals in a normal plot to assess the normal distributional assumption. A nonlinear association in such a q-q plot indicates nonnormality. If data-rich, a histogram of residuals can also be used.

Normal plots of residuals

1. Obtain the observed quantiles by ordering the residuals:
$$e_{(1)} \leq e_{(2)} \leq \cdots \leq e_{(n)}.$$
2. For each $i = 1, \ldots, n$ compute the expected quantile from
$$q_{(i)} = z(1 - \frac{i}{n+1}).$$
3. Plot the (ordered) residuals on the vertical axis versus the (ordered) theoretical quantiles on the horizontal axis.

The *empirical cumulative probability* associated with $e_{(i)}$ is

$$p_{(i)} = \frac{\text{Rank of } e_{(i)}}{n+1}.$$

Corresponding theoretical quantiles obtained via

$$q_{(i)} = z(1 - p_{(i)}).$$

e.g. suppose $n = 9$ then for $i = 1$ we look up the $10^{th}$ percentile of $N(0, 1)$ which is -1.282 ... for $i = 9$ we look up $q_{(9)} = +1.282$. Plot the ordered residuals (empirical quantiles) against the theoretical quantiles and expect linearity.

```
ods listing close;
ods graphics on;
proc reg data=running;
   model pace=sexf age age2;   *general linear model.  will discuss soon;
   output out=resids p=yhat r=resid;
run;
proc rank data=resids out=resids2;
   ranks rankresid;
   var resid;
run;
data resids2;
   set resids2;
   ecdf=rankresid/(160+1);    *160 runners;
   q=probit(ecdf);
run;
ods listing ;
proc print data=resids2 ;
   var age pace yhat resid rankresid ecdf q;
run;
proc gplot data=resids;
   plot resid*q;
run;
```

```
                          The SAS System                              1

 Obs    age      pace       yhat       resid     rankresid      ecdf         q

   1     28    5.3833     7.5837    -2.20040       14.0      0.08696    -1.35974
   2     39    5.4667     7.7671    -2.30046       10.0      0.06211    -1.53728
   3     41    5.5167     7.8735    -2.35681        6.0      0.03727    -1.78332
   4     42    5.6167     7.9351    -2.31841        9.0      0.05590    -1.59015
   5     40    5.9333     7.8175    -1.88416       18.0      0.11180    -1.21700
                      (abbreviated)
 156      6   14.4667    11.4534     3.01324      155.0      0.96273     1.78332
 157     52   15.1000    11.0579     4.04215      157.0      0.97516     1.96263
 158     10   17.2667    10.9473     6.31937      158.5      0.98447     2.15636
 159     10   17.2667    10.9473     6.31937      158.5      0.98447     2.15636
 160     81   17.5000    14.7178     2.78223      152.0      0.94410     1.59015
```

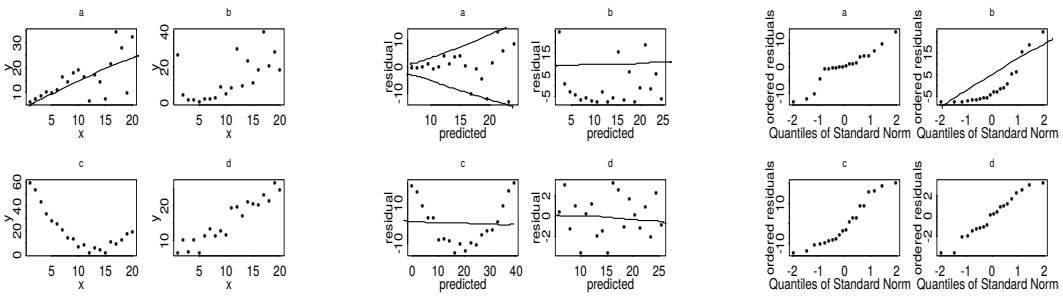
## A fun exercise: Match up letters a,b,c,d with the model violation

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

$\sigma^2(x) = Var(y|x) \propto x$

$e_i = y_i - \hat{y}_i$

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$




$y$ v $x$ resid v predicted ($y$ v $\hat{y}$) normal plots of resids

1. Heteroscedasticity (nonconstant Variance) (a) (c)
2. Nonlinearity ($\mu(x)$ not linear in $x$)
3. Nonnormality (vertical variation in $y$ about $\mu(x)$ not bell-shaped) (b)
4. Model fits (hurray!) (d)