Regression Review - Part II
Weighted Least Squares

Statistics 149

Spring 2006

Diagnostic Example

Body Fat Prediction

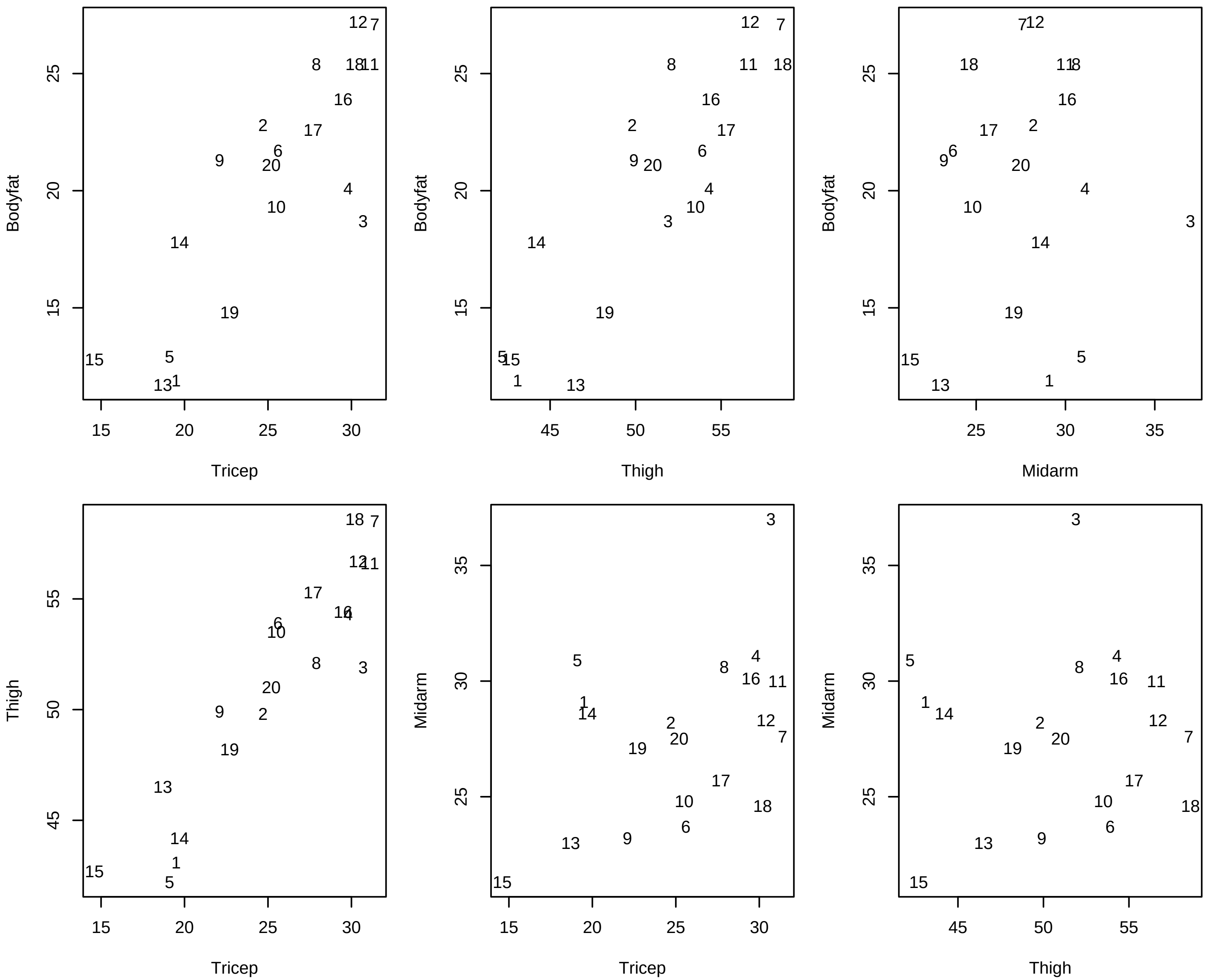
20 healthy females 25-34 years old were studied to come up with a predictive model for body fat based on simple measurements as the gold standard measurement is time consuming and expensive.

Response variable: `Bodyfat` - determined by body immersion in water

Predictor variables:

- `Tricep` - Triceps Skinfold Thickness
- `Thigh` - Thigh Circumference
- `Midarm` - Midarm Circumference

In the following figure, the plotting symbol is the observation number.

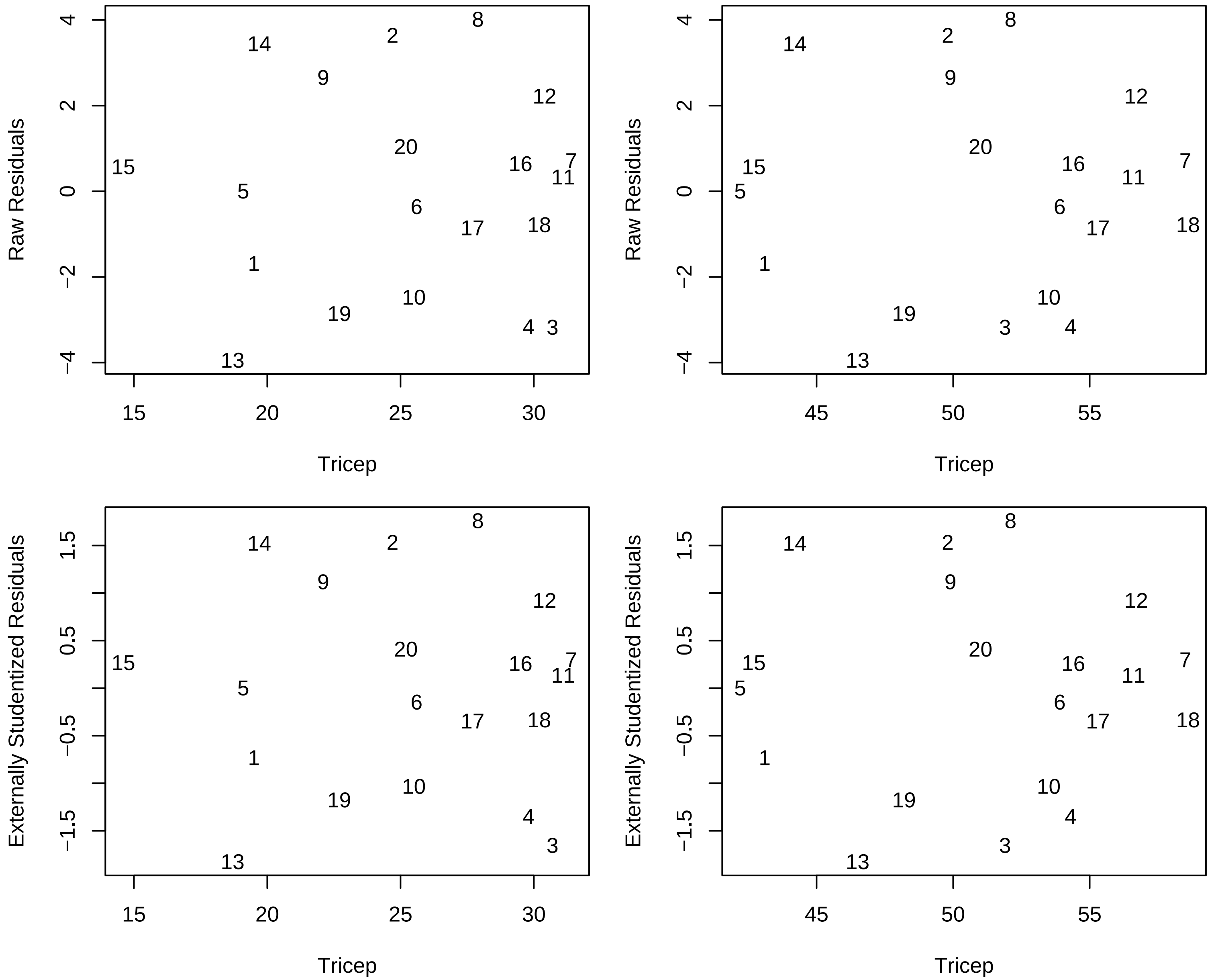

Bodyfat
Tricep
Thigh
Midarm

```
> cor(bodyfat)
        Tricep   Thigh  Midarm Bodyfat
Tricep  1.0000 0.92384 0.45778  0.8433
Thigh   0.9238 1.00000 0.08467  0.8781
Midarm  0.4578 0.08467 1.00000  0.1424
Bodyfat 0.8433 0.87809 0.14244  1.0000
```

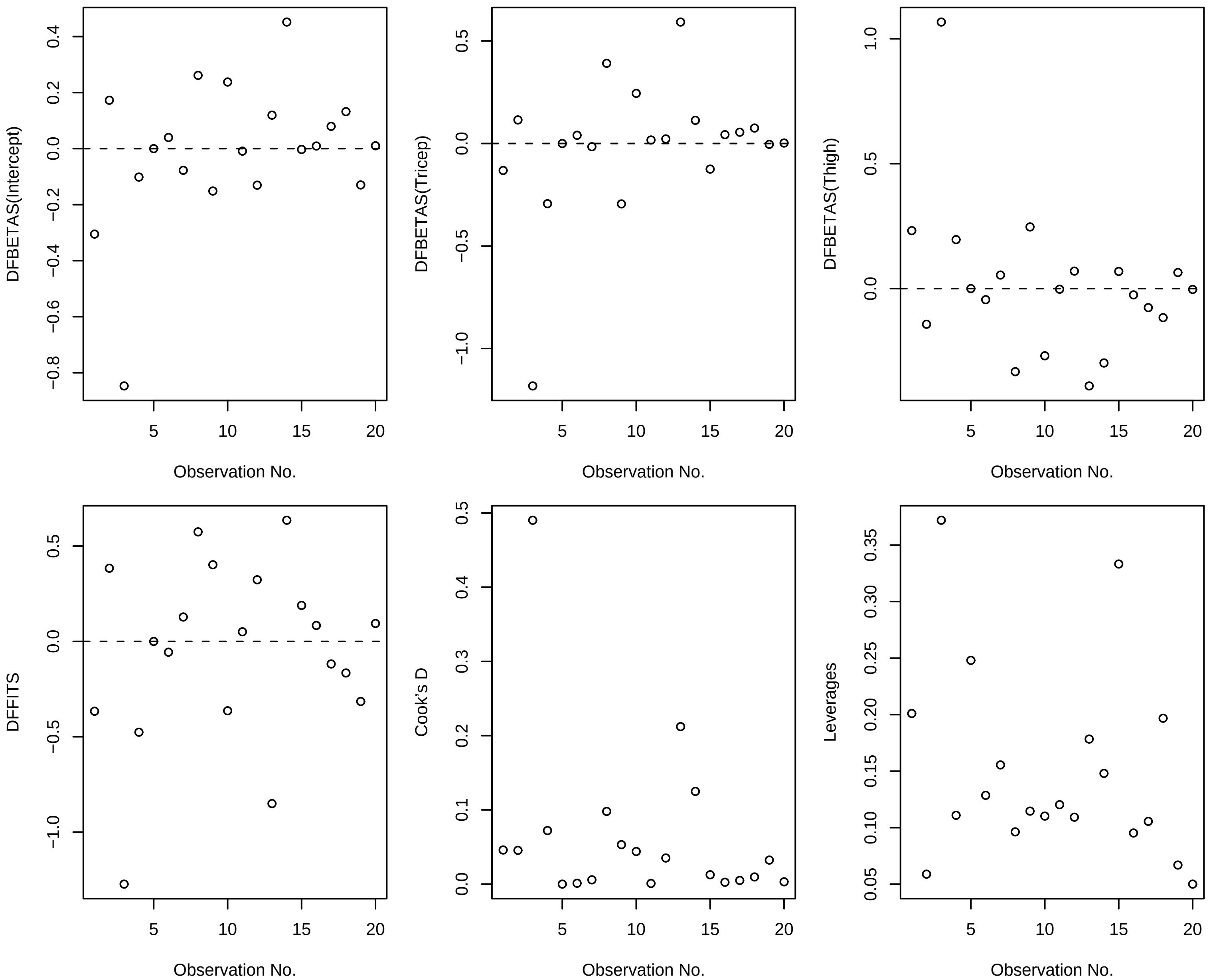
Lets fit the the model with `Tricep` and `Thigh` used to predict `Bodyfat`.

```
> bodyfat2.lm <- lm(Bodyfat ~ Tricep + Thigh, data=bodyfat)
>
> summary(bodyfat2.lm)

Call:
lm(formula = Bodyfat ~ Tricep + Thigh, data = bodyfat)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9469 -1.8807  0.1678  1.3367  4.0147
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.1742     8.3606  -2.293   0.0348 *
Tricep        0.2224     0.3034   0.733   0.4737
Thigh         0.6594     0.2912   2.265   0.0369 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.543 on 17 degrees of freedom
Multiple R-Squared: 0.7781,     Adjusted R-squared: 0.7519
F-statistic:  29.8 on 2 and 17 DF,  p-value: 2.774e-06
```



Raw Residuals
Tricep
Raw Residuals
Tricep
Externally Studentized Residuals
Tricep
Externally Studentized Residuals
Tricep


DFBETAS(Intercept)
DFBETAS(Tricep)
DFBETAS(Thigh)
DFFITS
Cook's D
Leverages
Observation No.

```
>Influence measures of
         lm(formula = Bodyfat ~ Tricep + Thigh, data = bodyfat) :

      dfb.1_  dfb.Trcp  dfb.Thgh     dffit cov.r   cook.d    hat inf
1  -3.05e-01 -1.31e-01  2.32e-01 -3.66e-01 1.361 4.60e-02 0.2010
2   1.73e-01  1.15e-01 -1.43e-01  3.84e-01 0.844 4.55e-02 0.0589
3  -8.47e-01 -1.18e+00  1.07e+00 -1.27e+00 1.189 4.90e-01 0.3719   *
4  -1.02e-01 -2.94e-01  1.96e-01 -4.76e-01 0.977 7.22e-02 0.1109
5  -6.37e-05 -3.05e-05  5.02e-05 -7.29e-05 1.595 1.88e-09 0.2480   *
6   3.97e-02  4.01e-02 -4.43e-02 -5.67e-02 1.371 1.14e-03 0.1286
7  -7.75e-02 -1.56e-02  5.43e-02  1.28e-01 1.397 5.76e-03 0.1555
8   2.61e-01  3.91e-01 -3.32e-01  5.75e-01 0.780 9.79e-02 0.0963
9  -1.51e-01 -2.95e-01  2.47e-01  4.02e-01 1.081 5.31e-02 0.1146
10  2.38e-01  2.45e-01 -2.69e-01 -3.64e-01 1.110 4.40e-02 0.1102
11 -9.02e-03  1.71e-02 -2.48e-03  5.05e-02 1.359 9.04e-04 0.1203
12 -1.30e-01  2.25e-02  7.00e-02  3.23e-01 1.152 3.52e-02 0.1093
13  1.19e-01  5.92e-01 -3.89e-01 -8.51e-01 0.827 2.12e-01 0.1784
14  4.52e-01  1.13e-01 -2.98e-01  6.36e-01 0.937 1.25e-01 0.1480
15 -3.00e-03 -1.25e-01  6.88e-02  1.89e-01 1.775 1.26e-02 0.3332   *
16  9.31e-03  4.31e-02 -2.51e-02  8.38e-02 1.309 2.47e-03 0.0953
17  7.95e-02  5.50e-02 -7.61e-02 -1.18e-01 1.312 4.93e-03 0.1056
```

```
18  1.32e-01  7.53e-02 -1.16e-01 -1.66e-01 1.462 9.64e-03 0.1968
19 -1.30e-01 -4.07e-03  6.44e-02 -3.15e-01 1.002 3.24e-02 0.0670
20  1.02e-02  2.29e-03 -3.31e-03  9.40e-02 1.224 3.10e-03 0.0501
```

So it appears that there is one obvious influential point (observations 3). Lets look at the parameter estimates in the two cases

	Intercept	Tricep	Thigh
All data	-19.17	0.22	0.66
Obs 3 dropped	-12.43	0.56	0.36

This particular observation is interesting in that this particular observation seems to have a smaller thigh measurement than would be expected given the the tricep measurement. In addition, this person has a midarm circumference at least 5 larger than anybody in the dataset.

The other two observations flagged by **R** don't seem to be particularly influential, especially observation 5. Both appear to be flagged because of the `covratio`).

R functions for diagnostics

- `resid(lm.object)`: raw residuals
- `fitted(lm.object)`: fitted values
- `rstandard(lm.object)`: studentized residuals
- `rstudent(lm.object)`: externally studentized residuals, aka deleted t residuals
- `dffits(lm.object)`

  R calls this influential if

$$DFFITS_i > 3\sqrt{\frac{(p+1)}{n-p-1}}$$

- `dfbetas(lm.object)`

  R calls this influential if

$$DFBETAS_{k(i)} > 1$$

- `cooks.distance(lm.object)`

  R calls this influential if

$$D_i > \text{The median of a } F_{p+1,n-p-1} \text{ distribution}$$

  This comes from origin of Cook's D, which was based on confidence ellipsoids for $\boldsymbol{\beta}$. These confidence ellipsoids involve F distributions.

- `hatvalues(lm.object)`

  R calls this influential if

$$h_i > \frac{3(p+1)}{n}$$

- `covratio(lm.object)`: This looks at the effect that the observation has on the estimate of σ. What is reported is

$$COVR_i = \left(\frac{\hat{\sigma}^2_{(i)}}{\hat{\sigma}^2}\right)^{p+1} \frac{1}{1-h_i}$$

  R calls this influential if

$$|1 - COVR_i| > \frac{3(p+1)}{n-p-1}$$

- `influence.measures(lm.object)`: gives previous 5 measures in the tabular format seen earlier.

Weighted Least Squares

One of the usual regression assumptions is that $\mathrm{Var}(Y_i|\mathbf{X}_i)$ is the same for all observations. However there may be situations where this isn't the case. Instead, the case maybe

$$\mathrm{Var}(Y_i|\mathbf{X}_i) = \frac{\sigma^2}{w_i}$$

where the w_i are known constants (known as weights).

Possible situations where this might hold are

- Responses are averages with known sample sizes:

$$\mathrm{Var}(Y_i|\mathbf{X}_i) = \frac{\sigma^2}{n_i}$$

- Responses are estimates and SEs are available:

  Sometimes the response variables are values are measurements whose estimated standard deviations $se(Y_i)$ are known. In this case,

$$w_i = \frac{1}{se(Y_i)^2}$$

- Variance is proportional to X (or a function of it):

  Sometimes while the regression of a response variable is a straight line, the variance increases with increases in the predictor variable. While a transformation on the response might solve the variance problem, it will introduce a nonlinear relationship. In this case

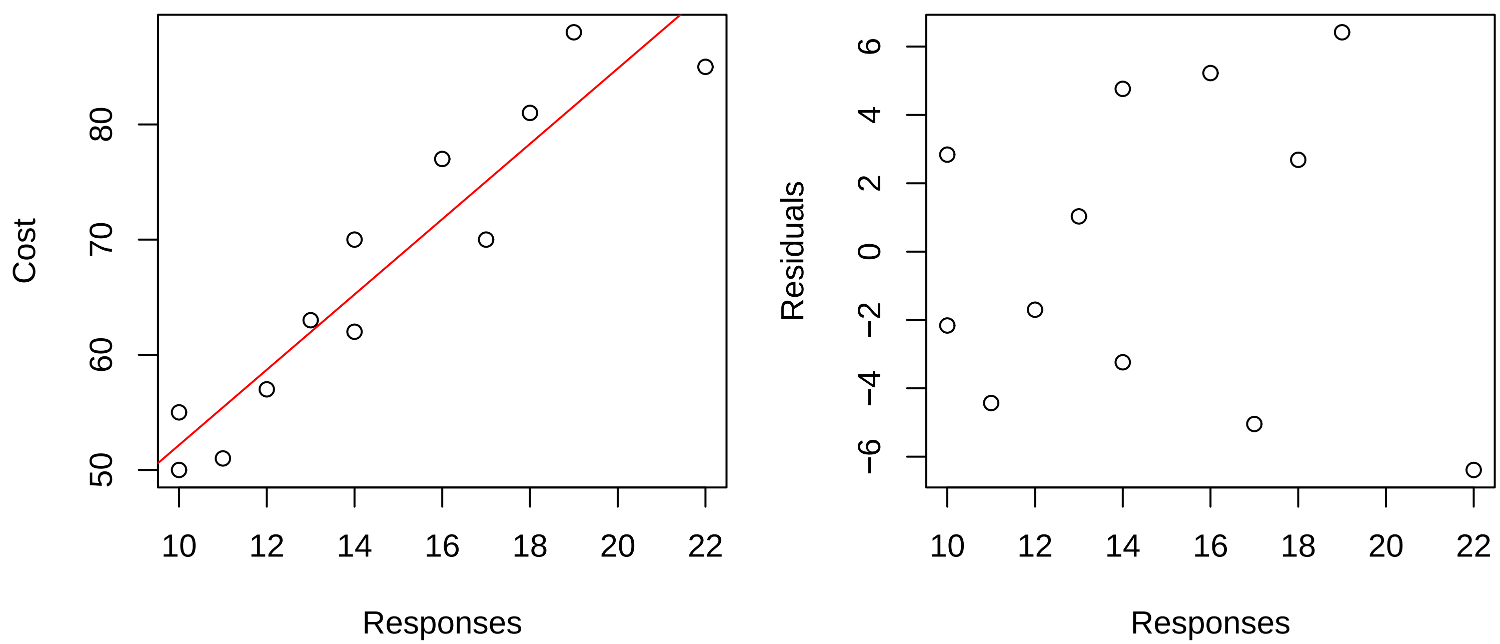
$$w_i = \frac{1}{X_i} \quad \text{or} \quad w_i = \frac{1}{X_i^2}$$

  might be reasonable.

Example: Computer-assisted learning

A study of computer-assisted learning in 12 students investigated the relationship between

- `Cost`: cost of computer time (in cents)
- `Responses`: the total number of responses in completing a lesson



So it appears that a linear relationship is reasonable, but the constant variance assumption isn't.

Instead it appears the standard deviation of the residuals might increase linearly with `Response`, or equivalently,

$$\sigma^2 \propto \texttt{Response}^2$$

This suggests an analysis with

$$w_i = \frac{1}{\texttt{Response}_i^2}$$

In this situation the least square estimate,

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

is still an unbiased estimate of $\boldsymbol{\beta}$ (see if you can show this), it is not minimum variance.

Intuitively this makes sense as if I know that $\mathrm{Var}(Y_i|\mathbf{X}_i)$ is small for certain observations, the regression surface should more likely be closer to these observations than ones with large $\mathrm{Var}(Y_i|\mathbf{X}_i)$.

So the idea behind weighted least squares is to weight observations with higher weights more. The weighted least squares criteria is

$$SS_w(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i(Y_i - \beta_0 - \beta_1 X_{i1} - \ldots - \beta_p X_{ip})^2$$

This penalizes big residuals for observations with big weights more that those with small residuals.

This can be written is a matrix formulation by defining the weight matrix

$$\mathbf{W} = \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_n \end{bmatrix}$$

(a diagonal matrix with the weights along the diagonal). Then

$$SS_w(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

The minimizer of this is given by the weighted least squares estimate

$$\hat{\boldsymbol{\beta}}_w = (\mathbf{X}\mathbf{W}\mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}$$

This is also a unbiased estimate of $\boldsymbol{\beta}$, but is has better variance properties than the least squares estimate.

The variance proportionality constant can be estimated by

$$\begin{aligned}
\hat{\sigma}_w^2 &= \frac{1}{n-p-1}\sum_{i=1}^{n} w_i (Y_i - \hat{Y}_i)^2 \\
&= \frac{1}{n-p-1}\sum_{i=1}^{n} w_i e_i^2 \\
&= \frac{(\mathbf{Y} - \hat{\mathbf{Y}})^T \mathbf{W} (\mathbf{Y} - \hat{\mathbf{Y}})}{n-p-1}
\end{aligned}$$

and the variance of $\hat{\boldsymbol{\beta}}_w$ is given by

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_w) = \sigma_w^2 (\mathbf{XWX})^{-1}$$

which is usually estimated by

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_w) = \hat{\sigma}_w^2 (\mathbf{XWX})^{-1}$$

For the example, the weighted least squares analysis gives

```
> learning.w.lm <- lm(Cost ~ Responses, data= learning,
    weight=1/(Responses^2))

> summary(learning.w.lm)

Call:
lm(formula = Cost ~ Responses, data = learning,
     weights = 1/(Responses^2))

Residuals:
     Min       1Q   Median       3Q      Max
-0.36027 -0.25080 -0.01040  0.30517  0.34470
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.4530     4.8970   3.564  0.00515 **
Responses     3.4100     0.3649   9.346 2.94e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2975 on 10 degrees of freedom
Multiple R-Squared: 0.8973,     Adjusted R-squared: 0.887
F-statistic: 87.34 on 1 and 10 DF,  p-value: 2.945e-06

> vcov(learning.w.lm)
            (Intercept) Responses
(Intercept)       23.98    -1.737
Responses         -1.74     0.133
```

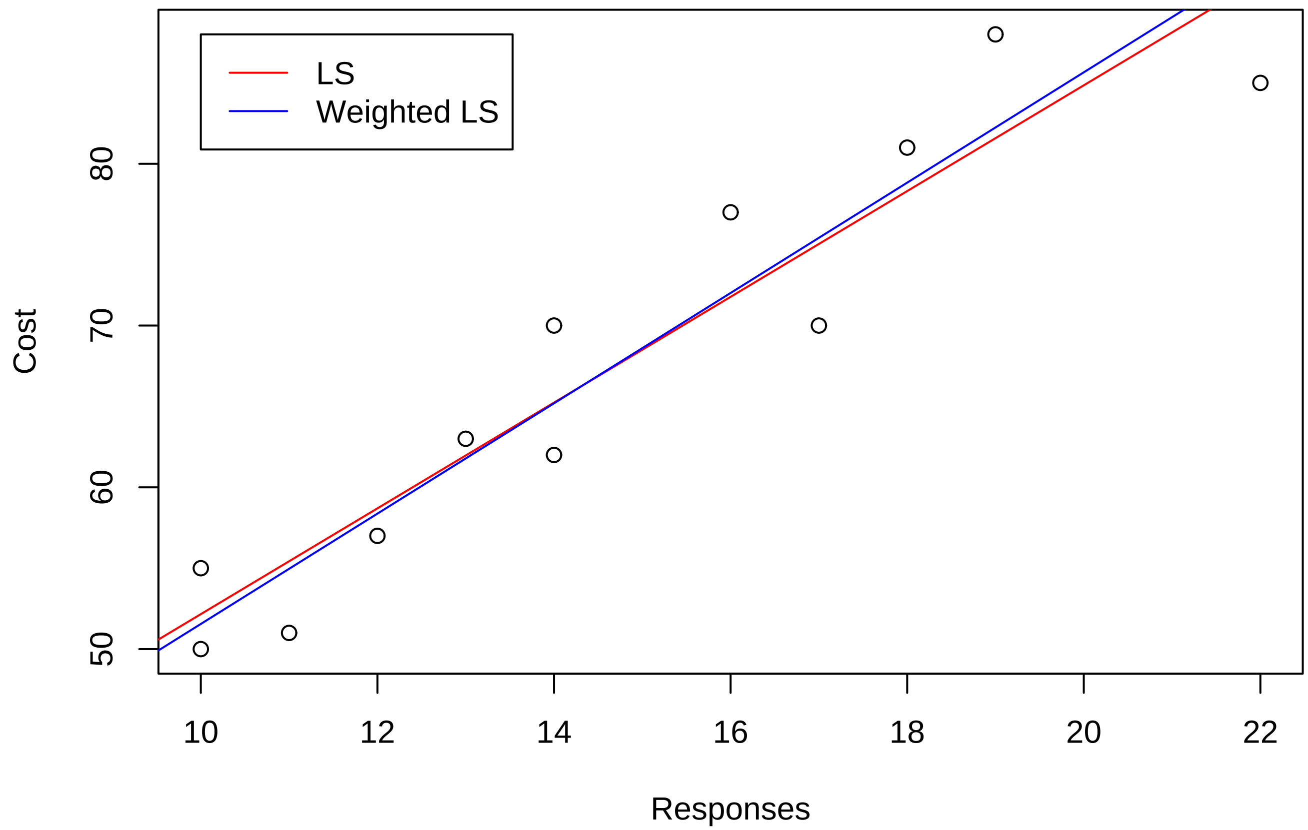
In this output, $\hat{\sigma}_w^2$ is given by `Residual standard error: 0.2975`

The `vcov` line gives $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}_w)$ the usual estimate of $\mathrm{Var}(\hat{\boldsymbol{\beta}}_w)$.

For comparison, here is the regular least squares analysis.

```
> summary(learning.lm)

Call:
lm(formula = Cost ~ Responses, data = learning)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3887 -3.5357 -0.3340  3.3193  6.4181

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.4727     5.5162   3.530  0.00545 **
Responses     3.2689     0.3651   8.955 4.33e-06 ***

Residual standard error: 4.598 on 10 degrees of freedom
Multiple R-Squared: 0.8891,     Adjusted R-squared: 0.878
F-statistic: 80.19 on 1 and 10 DF,  p-value: 4.33e-06
```

```
> vcov(learning.lm)
            (Intercept) Responses
(Intercept)       30.43    -1.955
Responses         -1.95     0.133
```



Note that the residual summaries from both analyzes are quite different

From WLS analysis

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.36027 -0.25080 -0.01040  0.30517  0.34470
```

From LS analysis

```
Residuals:
    Min      1Q  Median      3Q     Max
-6.3887 -3.5357 -0.3340  3.3193  6.4181
```

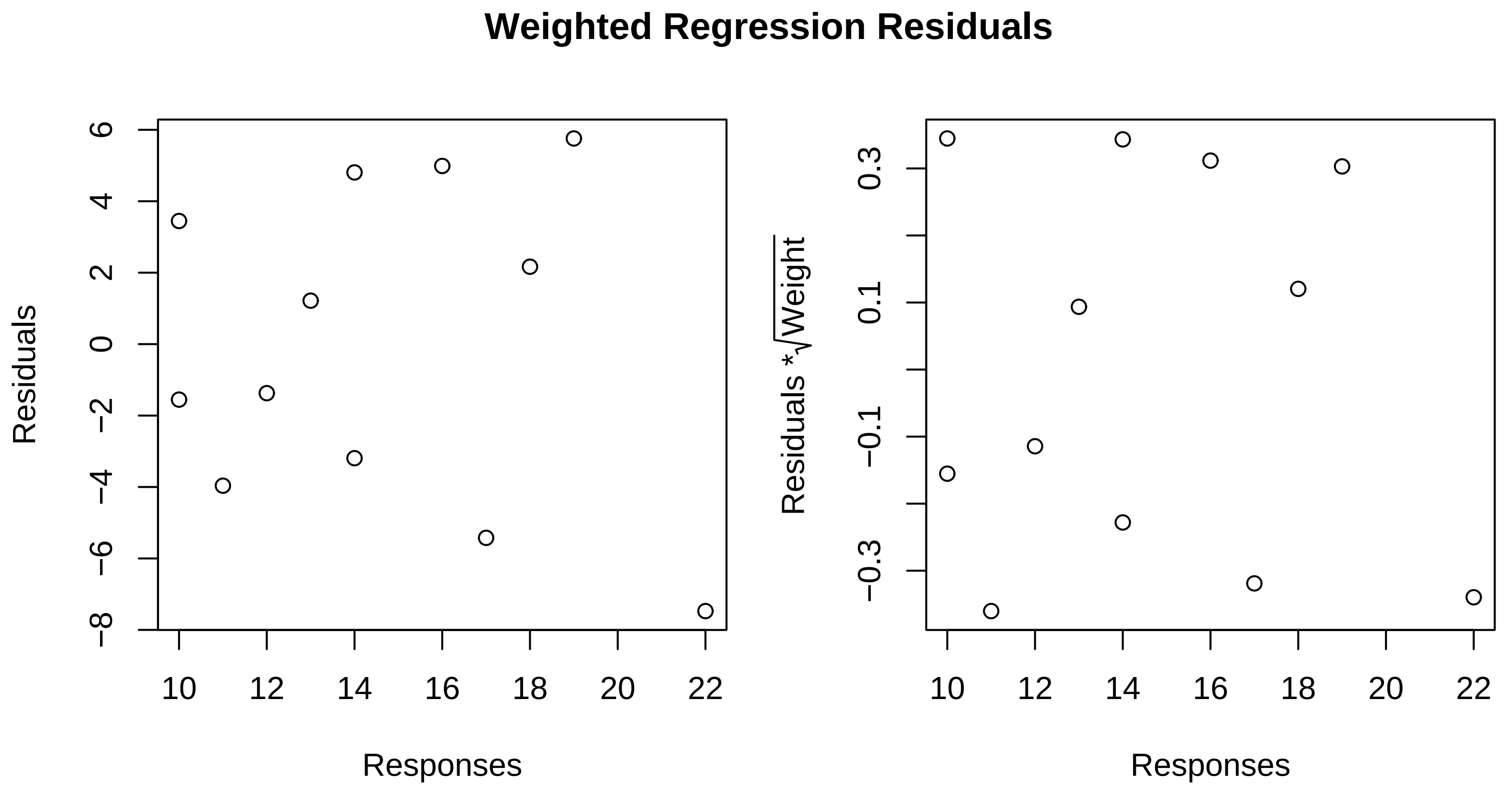
This is due to the different assumptions about the variances. If

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

then

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

In the residual summary of the weighted least squares analysis, this is based on $e_i\sqrt{w_i}$ instead of the raw residuals e_i. In addition, to see if a reasonable weighting has been done, plot $e_i\sqrt{w_i}$ instead of e_i



In this case, using weights of $\frac{1}{\texttt{Response}_i^2}$ seams reasonable.