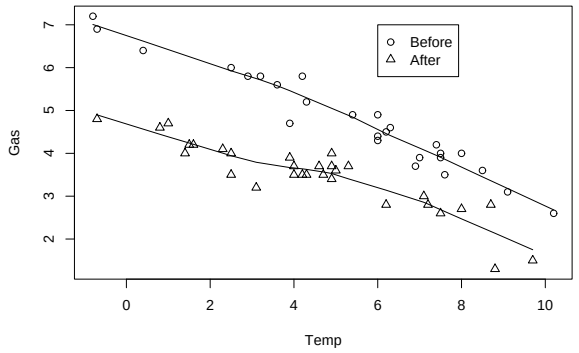
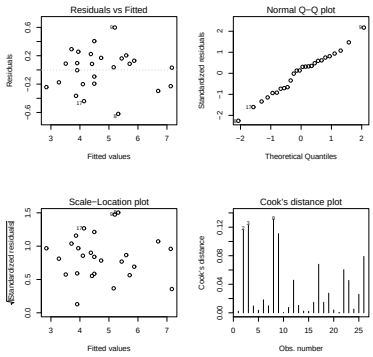
Gas Consumption

```
> library(MASS)
> names(whiteside)
[1] "Insul" "Temp"  "Gas"
> attach(whiteside)
> table(Insul)
Insul
Before  After
    26     30
> plot(Gas ~ Temp, pch=as.numeric(Insul))
> legend(6,7,c("Before","After"),pch=c(1,2))
> lines(lowess(Temp[Insul=="Before"],Gas[Insul=="Before"]))
> lines(lowess(Temp[Insul=="After"],Gas[Insul=="After"]))
>
```



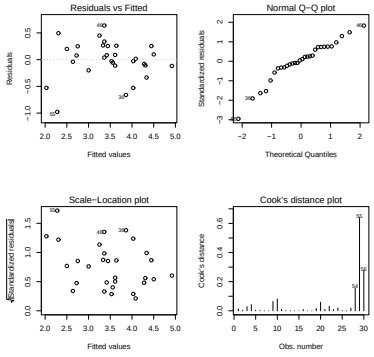
Simple Linear Model: Before

```
> lb <- lm(Gas ~ Temp, data=whiteside, subset=Insul=="Before")
> summary(lb)
Call:
lm(formula = Gas ~ Temp, data = whiteside, subset = Insul ==
    "Before")
[ . . . ]
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.85383    0.11842   57.88   <2e-16 ***
Temp        -0.39324    0.01959  -20.08   <2e-16 ***
[ . . . ]
Residual standard error: 0.2813 on 24 degrees of freedom
Multiple R-Squared: 0.9438,Adjusted R-squared: 0.9415
F-statistic: 403.1 on 1 and 24 DF,  p-value: < 2.2e-16
> plot(lb)
```



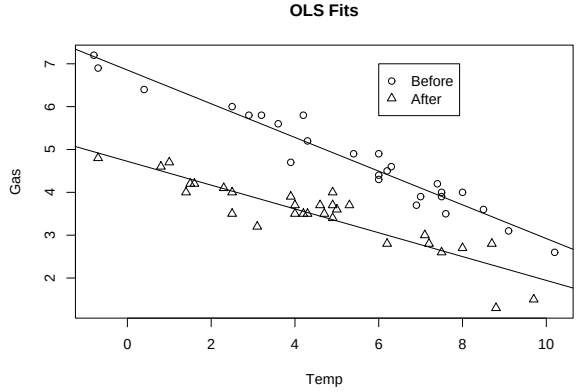
Simple Linear Model: After

```
> la <- update(lb, subset=Insul=="After")
> summary(la)
Call:
lm(formula = Gas ~ Temp, data = whiteside, subset = Insul ==
    "After")
[ . . . ]
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.72385    0.12974   36.41  < 2e-16 ***
Temp        -0.27793    0.02518  -11.04 1.05e-11 ***
[ . . . ]
Residual standard error: 0.3548 on 28 degrees of freedom
Multiple R-Squared: 0.8131,Adjusted R-squared: 0.8064
F-statistic: 121.8 on 1 and 28 DF,  p-value: 1.046e-11
> plot(la)
```



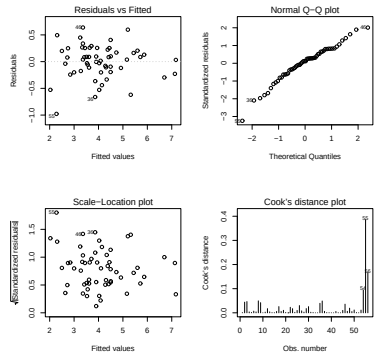
Comparison: Before and After

```
> coef(summary(la))
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  4.7238497 0.12973942  36.41029 3.958906e-25
Temp        -0.2779350 0.02518429 -11.03605 1.045745e-11
> coef(summary(lb))
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  6.8538277 0.11842341  57.87561 2.717533e-27
Temp        -0.3932388 0.01958601 -20.07754 1.640469e-16
> plot(Gas ~ Temp, pch=as.numeric(Insul), main="OLS Fits")
> legend(6,7,c("Before","After"),pch=c(1,2))
> abline(la); abline(lb)
```



Combined Model

```
> lcomb <- lm(Gas ~ Insul/Temp - 1, data=whiteside)
> summary(lcomb)
Call:
lm(formula = Gas ~ Insul/Temp - 1, data = whiteside)
[ . . . ]
                 Estimate Std. Error t value Pr(>|t|)
InsulBefore       6.85383    0.13596   50.41   <2e-16 ***
InsulAfter        4.72385    0.11810   40.00   <2e-16 ***
InsulBefore:Temp -0.39324    0.02249  -17.49   <2e-16 ***
InsulAfter:Temp  -0.27793    0.02292  -12.12   <2e-16 ***
[ . . . ]
Residual standard error: 0.323 on 52 degrees of freedom
Multiple R-Squared: 0.9946,Adjusted R-squared: 0.9942
F-statistic:  2391 on 4 and 52 DF,  p-value: < 2.2e-16
> plot(lcomb)
```



Comparison

```
> summary(lb)
[ . . . ]
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.85383    0.11842   57.88   <2e-16 ***
Temp        -0.39324    0.01959  -20.08   <2e-16 ***
[ . . . ]
Residual standard error: 0.2813 on 24 degrees of freedom
Multiple R-Squared: 0.9438,Adjusted R-squared: 0.9415
F-statistic: 403.1 on 1 and 24 DF,  p-value: < 2.2e-16
> summary(la)
[ . . . ]
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.72385    0.12974   36.41  < 2e-16 ***
Temp        -0.27793    0.02518  -11.04 1.05e-11 ***
[ . . . ]
Residual standard error: 0.3548 on 28 degrees of freedom
Multiple R-Squared: 0.8131,Adjusted R-squared: 0.8064
F-statistic: 121.8 on 1 and 28 DF,  p-value: 1.046e-11
> summary(lcomb)
[ . . . ]
                 Estimate Std. Error t value Pr(>|t|)
InsulBefore       6.85383    0.13596   50.41   <2e-16 ***
InsulAfter        4.72385    0.11810   40.00   <2e-16 ***
InsulBefore:Temp -0.39324    0.02249  -17.49   <2e-16 ***
InsulAfter:Temp  -0.27793    0.02292  -12.12   <2e-16 ***
[ . . . ]
Residual standard error: 0.323 on 52 degrees of freedom
Multiple R-Squared: 0.9946,Adjusted R-squared: 0.9942
F-statistic:  2391 on 4 and 52 DF,  p-value: < 2.2e-16
>
```


Analysis of Covariance

For $k \in \{1, 2\}$ corresponding to the "Before" and "After" groups, can the model

$$\mathsf{Gas}_{k,j} = \alpha_k + \beta_k \mathsf{Temp}_{k,j} + \epsilon_{k,j}$$

be replaced by the following simpler model?

$$\mathsf{Gas}_{k,j} = \alpha_k + \beta \mathsf{Temp}_{k,j} + \epsilon_{k,j}$$

We can fit the latter model by:

```
> lpara <- lm(Gas ~ Insul + Temp - 1, data=whiteside)
> summary(lpara)
[ . . . ]
           Estimate Std. Error t value Pr(>|t|)
InsulBefore  6.55133    0.11809   55.48   <2e-16 ***
InsulAfter   4.98612    0.10268   48.56   <2e-16 ***
Temp        -0.33670    0.01776  -18.95   <2e-16 ***
[ . . . ]
Residual standard error: 0.3574 on 53 degrees of freedom
Multiple R-Squared: 0.9933,Adjusted R-squared: 0.9929
F-statistic:  2600 on 3 and 53 DF,  p-value: < 2.2e-16
>
```

and compare two such "nested" models with:

```
> anova(lpara, lcomb)
Analysis of Variance Table

Model 1: Gas ~ Insul + Temp - 1
Model 2: Gas ~ Insul/Temp - 1
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     53 6.7704
2     52 5.4252  1    1.3451 12.893 0.0007307 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```


An Alternate Parametrization

The alternative parametrization fit by:

```
> summary(lmalt <- lm(Gas ~ Insul*Temp, data=whiteside))
[ . . . ]
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.85383    0.13596  50.409  < 2e-16 ***
InsulAfter   -2.12998    0.18009 -11.827 2.32e-16 ***
Temp         -0.39324    0.02249 -17.487  < 2e-16 ***
InsulAfter:Temp  0.11530    0.03211   3.591  0.00073 ***
[ . . . ]
Residual standard error: 0.323 on 52 degrees of freedom
Multiple R-Squared: 0.9277,Adjusted R-squared: 0.9235
F-statistic: 222.3 on 3 and 52 DF,  p-value: < 2.2e-16
>
```

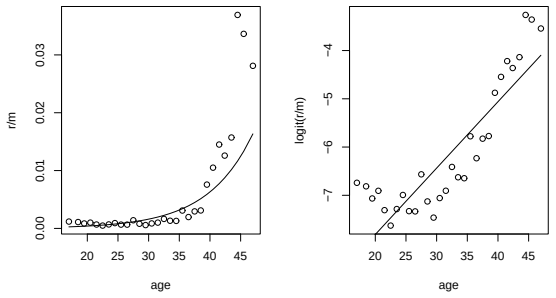
corresponds to the model:

$$\mathsf{Gas}_j = \alpha_{\mathsf{Before}} + \alpha_{\mathsf{diff}} I_{\mathsf{After},j} + \beta_{\mathsf{Before}} \mathsf{Temp}_j + \beta_{\mathsf{diff}} I_{\mathsf{After},j} \mathsf{Temp}_j$$

where $I_{\mathsf{After},j}$ is an indicator for "after insulation". Note that the slope for "before insulation" is β_{Before}, and the slope for "after insulation" is $\beta_{\mathsf{Before}} + \beta_{\mathsf{diff}}$, so the fourth t-test corresponding to $H_0\colon \beta_{\mathsf{diff}} = 0$ is a test of no difference in slopes before and after insulation.

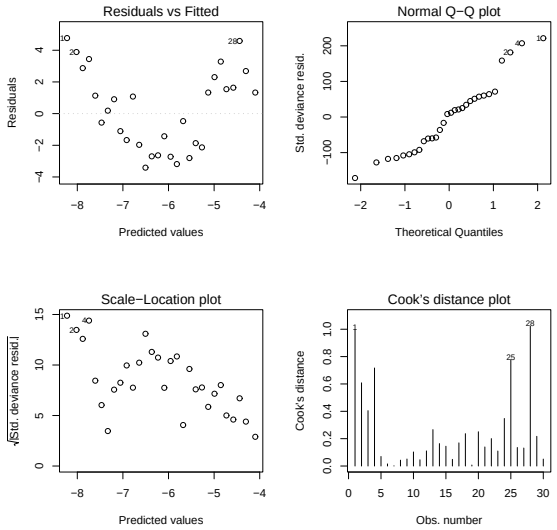
Down's Syndrome Data

```
> library(boot)
> attach(downs.bc)
> downs.glm1 <- glm(cbind(r,m-r) ~ age, family=binomial, data=downs.bc)
> plot(age, r/m)
> lines(age, fitted(downs.glm1))
> plot(age, logit(r/m))
> lines(age, predict(downs.glm1))
```



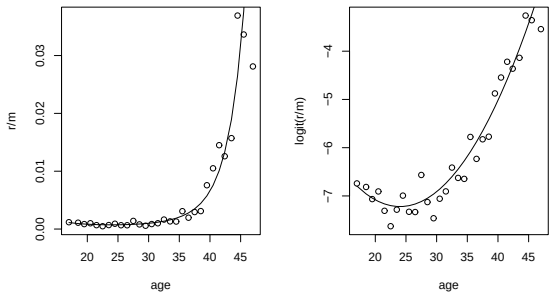
Linear Fit

```
> plot(downs.glm1)
```



Quadratic Fit

```
> downs.glm2 <- glm(cbind(r,m-r) ~ age + I(age^2), family=binomial,
+ data=downs.bc)
> plot(age,r/m)
> lines(age,fitted(downs.glm2))
> plot(age, logit(r/m))
> lines(age, predict(downs.glm2))
```




```
> coef(summary(downs.glm2))
               Estimate   Std. Error   z value     Pr(>|z|)
(Intercept) -2.243581610 0.6634504391 -3.381687 7.204223e-04
age         -0.414359388 0.0436091347 -9.501665 2.065608e-21
I(age^2)     0.008624321 0.0006780733 12.718862 4.645754e-37
>
```


Linear, Quadratic, Cubic?

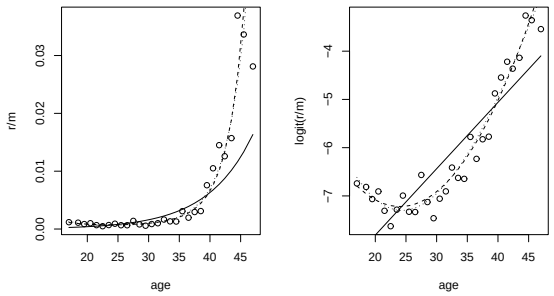
```
> downs.glm3 <- update(downs.glm2, . ~ . + I(age^3))
> anova(downs.glm1, downs.glm2, downs.glm3)
Analysis of Deviance Table

Model 1: cbind(r, m - r) ~ age
Model 2: cbind(r, m - r) ~ age + I(age^2)
Model 3: cbind(r, m - r) ~ age + I(age^2) + I(age^3)
  Resid. Df Resid. Dev Df Deviance
1        28    184.027
2        27     44.787  1  139.240
3        26     42.109  1    2.678
> anova(downs.glm1, downs.glm2, downs.glm3, test="Chisq")
Analysis of Deviance Table

Model 1: cbind(r, m - r) ~ age
Model 2: cbind(r, m - r) ~ age + I(age^2)
Model 3: cbind(r, m - r) ~ age + I(age^2) + I(age^3)
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        28    184.027
2        27     44.787  1  139.240 3.902e-32
3        26     42.109  1    2.678     0.102
>
```

Linear, Quadratic, Cubic?

```
> plot(age, r/m)
> lines(age, fitted(downs.glm1),lty=1)
> lines(age, fitted(downs.glm2),lty=2)
> lines(age, fitted(downs.glm3),lty=3)
> plot(age, logit(r/m))
> lines(age, predict(downs.glm1),lty=1)
> lines(age, predict(downs.glm2),lty=2)
> lines(age, predict(downs.glm3),lty=3)
```



Updating Models

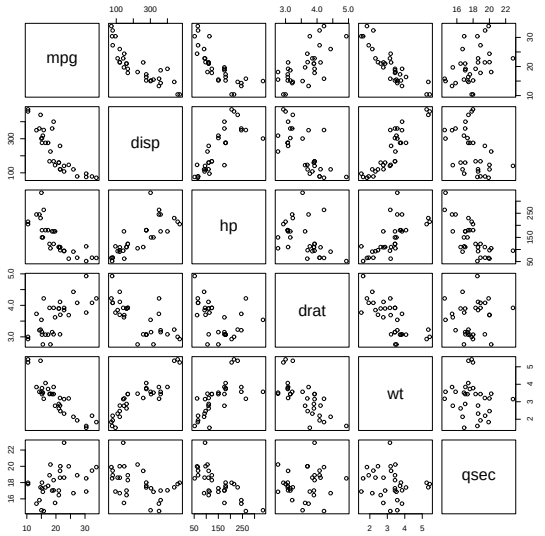
- `update(lm.fit, subset=...)` Fit the model to a different subset of the dataframe.
  - `update(lm.fit, subset=c(-10,-15))` Refit without observations 10 and 15.
  - `update(lm.fit, subset=dose==0)` Refit model to placebo group only.
- `update(lm.fit, .~.`$\pm terms$`)` Add or remove terms from the model.
  - `update(lm.fit, .~.+age-birth.year)` Add `age` and remove `birth.year` terms.
  - `update(lm.fit, . ~ sex/.-1)` Fit the model (whatever it was) separately for both sexes (and remove the common intercept).
- `update(lm.fit, f(.)~.)` Refit the model with transformed response.
  - `update(lm.fit, log(.) ~ .)` Use log-transformation for the response.
  - `update(lm.fit, 1/.^2 ~ .)` Replace the old response y with $1/y^2$.

Another `update` Example

```
> ll <- lm(Gas ~ Temp, data=whiteside)
> lb <- update(ll, subset=Insul=="Before")
> la <- update(ll, subset=Insul=="After")
> coef(lb)
(Intercept)        Temp
  6.8538277  -0.3932388
> coef(la)
(Intercept)        Temp
  4.7238497  -0.2779350
> lcomb <- update(ll, . ~ Insul/.)
> summary(lcomb)
[ . . . ]
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.85383    0.13596   50.41  < 2e-16 ***
InsulAfter       -2.12998    0.18009  -11.83 2.32e-16 ***
InsulBefore:Temp -0.39324    0.02249  -17.49  < 2e-16 ***
InsulAfter:Temp  -0.27793    0.02292  -12.12  < 2e-16 ***
[ . . . ]
> lcomb <- update(lcomb, . ~ . -1)
> summary(lcomb)
[ . . . ]
                Estimate Std. Error t value Pr(>|t|)
InsulBefore       6.85383    0.13596   50.41   <2e-16 ***
InsulAfter        4.72385    0.11810   40.00   <2e-16 ***
InsulBefore:Temp -0.39324    0.02249  -17.49   <2e-16 ***
InsulAfter:Temp  -0.27793    0.02292  -12.12   <2e-16 ***
[ . . . ]
```

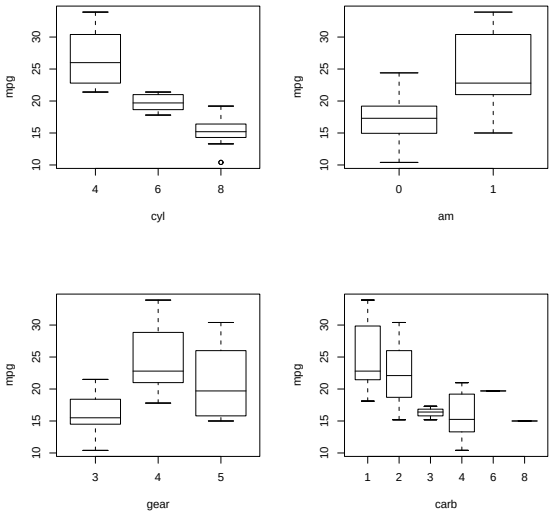

Motor Trend Car Tests

```
> names(mtcars)
 [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"
 [9] "am"   "gear" "carb"
> plot(mtcars[,c("mpg","disp","hp","drat","wt","qsec")])
```



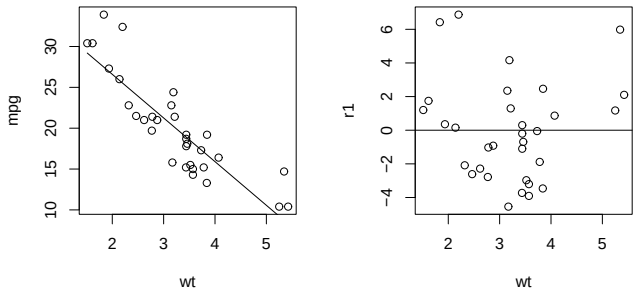
Motor Trend Car Tests

```
> for (i in c("cyl","am","gear","carb")) {
+   mtcars[,i] <- factor(mtcars[,i])
+ }
> attach(mtcars)
> plot(mpg ~ cyl)
> plot(mpg ~ am)
> plot(mpg ~ gear)
> plot(mpg ~ carb)
```



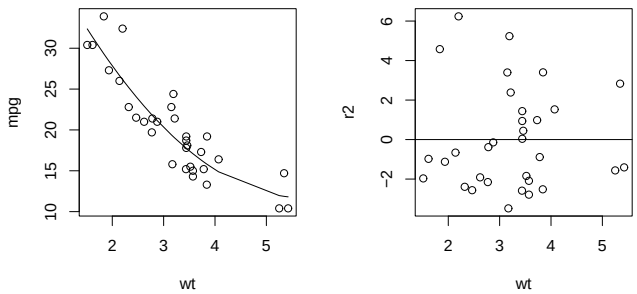
Regression on Weight

```
> cor(mtcars$mpg, mtcars[,c("disp","hp","drat","wt","qsec")])
          disp         hp      drat         wt     qsec
[1,] -0.8475514 -0.7761684 0.6811719 -0.8676594 0.418684
> l1 <- lm(mpg ~ wt, data=mtcars)
> summary(l1)
[ . . . ]
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2851     1.8776  19.858  < 2e-16 ***
wt           -5.3445     0.5591  -9.559 1.29e-10 ***
[ . . . ]
Residual standard error: 3.046 on 30 degrees of freedom
Multiple R-Squared: 0.7528,Adjusted R-squared: 0.7446
F-statistic: 91.38 on 1 and 30 DF,  p-value: 1.294e-10
> plot(mpg ~ wt)
> o <- order(wt); lines(wt[o],fitted(l1)[o])
> plot(resid(l1) ~ wt); abline(h=0)
```



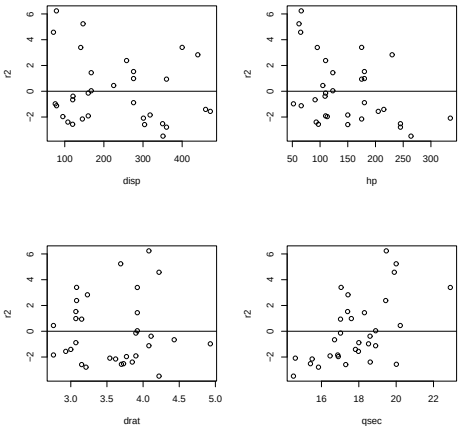
Quadratic Fit

```
> l2 <- update(l1, .~.+I(wt^2))
> anova(l1,l2)
[ . . . ]
Model 1: mpg ~ wt
Model 2: mpg ~ wt + I(wt^2)
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     30 278.322
2     29 203.745  1    74.576 10.615 0.00286 **
> plot(mpg ~ wt)
> o <- order(wt); lines(wt[o],fitted(l2)[o])
> plot(resid(l2) ~ wt); abline(h=0)
```



Residuals vs. Other Vars

```
> r2 <- resid(l2)
> plot(r2 ~ disp); abline(h=0)
> plot(r2 ~ hp); abline(h=0)
> plot(r2 ~ drat); abline(h=0)
> plot(r2 ~ qsec); abline(h=0)
```



One-Term Additions with `add1`

The `add1` function can be used to "grow" a model by one term within a given scope:

```
> add1(l2, ~.+disp+hp+drat+qsec, test="F")
Single term additions

Model:
mpg ~ wt + I(wt^2)
       Df Sum of Sq     RSS     AIC F value     Pr(F)
<none>              203.745  65.236
disp    1    30.705 173.040  62.009  4.9685 0.0340156 *
hp      1    59.452 144.293  56.196 11.5367 0.0020606 **
drat    1     0.276 203.470  67.193  0.0379 0.8470048
qsec    1    70.431 133.315  53.663 14.7925 0.0006339 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> 32*log(203.745/32)+2*3
[1] 65.23627
>
```

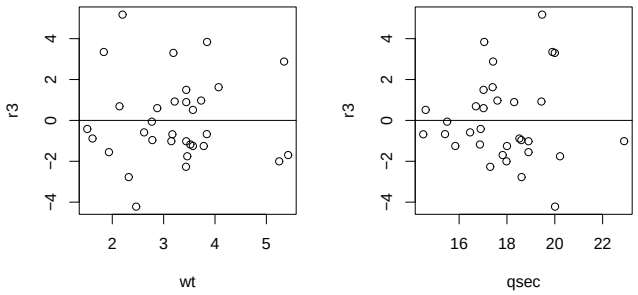
Besides comparing RSS values, you can look at the

$$\text{AIC} = n \log\left(\frac{\text{RSS}}{n}\right) + 2k$$

where n is number of obs and k is number of terms (including intercept) in the model. A lower AIC means a "better" model.

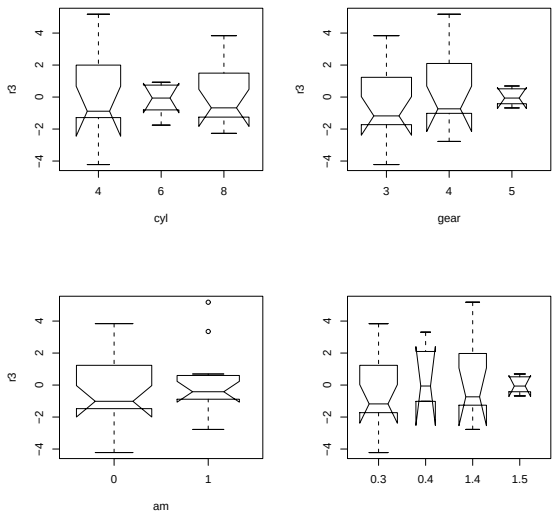
Weight and Quarter-Sec

```
> l3 <- update(l2,~.+qsec)
> summary(l3)
[ . . . ]
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.6418     5.6768   5.750 3.59e-06 ***
wt          -12.4331     2.0842  -5.965 2.01e-06 ***
I(wt^2)       1.0730     0.2970   3.613 0.001174 **
qsec          0.8599     0.2236   3.846 0.000634 ***
[ . . . ]
> anova(l3, update(l3, .~.+wt:qsec))
Model 1: mpg ~ wt + I(wt^2) + qsec
Model 2: mpg ~ wt + I(wt^2) + qsec + wt:qsec
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     28 133.315
2     27 128.033  1     5.282 1.1138 0.3006
>
```



Useful Factors?

```
> r3 <- resid(l3)
> plot(r3 ~ cyl, notch=T, varwidth=T)
> plot(r3 ~ gear, notch=T, varwidth=T)
> plot(r3 ~ am, notch=T, varwidth=T)
> boxplot(r3 ~ am*gear, notch=T, varwidth=T)
```



Useful Factors?

```
> anova(l3, update(l3, .~cyl/.))
Analysis of Variance Table

Model 1: mpg ~ wt + I(wt^2) + qsec
Model 2: mpg ~ cyl + cyl:wt + cyl:I(wt^2) + cyl:qsec
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     28 133.315
2     20 122.355  8    10.959 0.2239 0.9822
> anova(l3, update(l3, .~(am*gear)/.))
Analysis of Variance Table

Model 1: mpg ~ wt + I(wt^2) + qsec
Model 2: mpg ~ am + gear + am:gear + am:gear:wt
         + am:gear:I(wt^2) + am:gear:qsec
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1     28 133.315
2     16  86.693 12    46.622 0.717 0.7167
>
```

Quick Up/Down Search

Double-check that there are no "extra" terms that help:

```
> add1(l3, ~ .+disp+hp+drat+I(wt^3)+I(qsec^2), test="F")
[ . . . ]
mpg ~ wt + I(wt^2) + qsec
          Df Sum of Sq     RSS     AIC F value  Pr(F)
<none>                  133.315  53.663
disp       1     0.216 133.098  55.611  0.0439 0.8356
hp         1     2.722 130.593  55.003  0.5628 0.4596
drat       1     1.374 131.941  55.332  0.2811 0.6003
I(wt^3)    1     1.805 131.510  55.227  0.3706 0.5477
I(qsec^2)  1     0.913 132.402  55.443  0.1862 0.6695
```

Double-check that there are no "useless" terms still in the model:

```
> drop1(l3, test="F")
[ . . . ]
mpg ~ wt + I(wt^2) + qsec
        Df Sum of Sq     RSS     AIC F value     Pr(F)
<none>              133.315  53.663
wt       1   169.437 302.752  77.910  35.587 2.008e-06 ***
I(wt^2)  1    62.149 195.464  63.908  13.053 0.0011739 **
qsec     1    70.431 203.745  65.236  14.793 0.0006339 ***
```


"Final" Model

```
> summary(l3)
lm(formula = mpg ~ wt + I(wt^2) + qsec, data = mtcars)
[ . . . ]
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.6418     5.6768   5.750 3.59e-06 ***
wt          -12.4331     2.0842  -5.965 2.01e-06 ***
I(wt^2)       1.0730     0.2970   3.613 0.001174 **
qsec          0.8599     0.2236   3.846 0.000634 ***
[ . . . ]
Residual standard error: 2.182 on 28 degrees of freedom
Multiple R-Squared: 0.8816,Adjusted R-squared: 0.8689
F-statistic:  69.5 on 3 and 28 DF,  p-value: 4.345e-13
> plot(l3)
```