MLR: Model Checking

Math 430, Winter

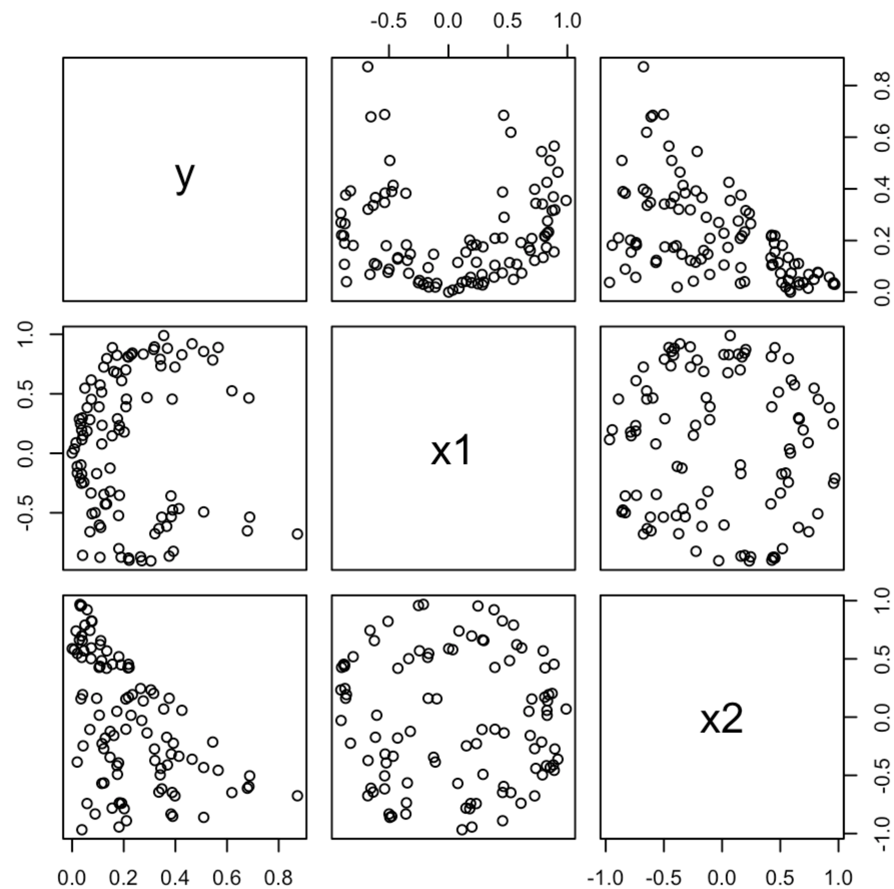
Toolkit

For model diagnostics, the `car` package is particularly useful:

```
library(car)
```

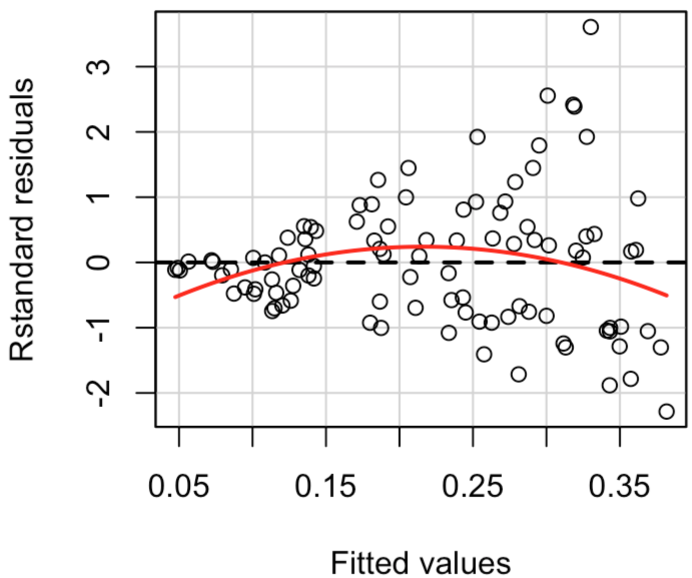
Assessing model assumptions

CAUTION: Residuals in MLR



CAUTION: Residuals in MLR

When there are many regressors in a model, we cannot necessarily associate shapes in a residual plot with a particular problem with the assumptions.



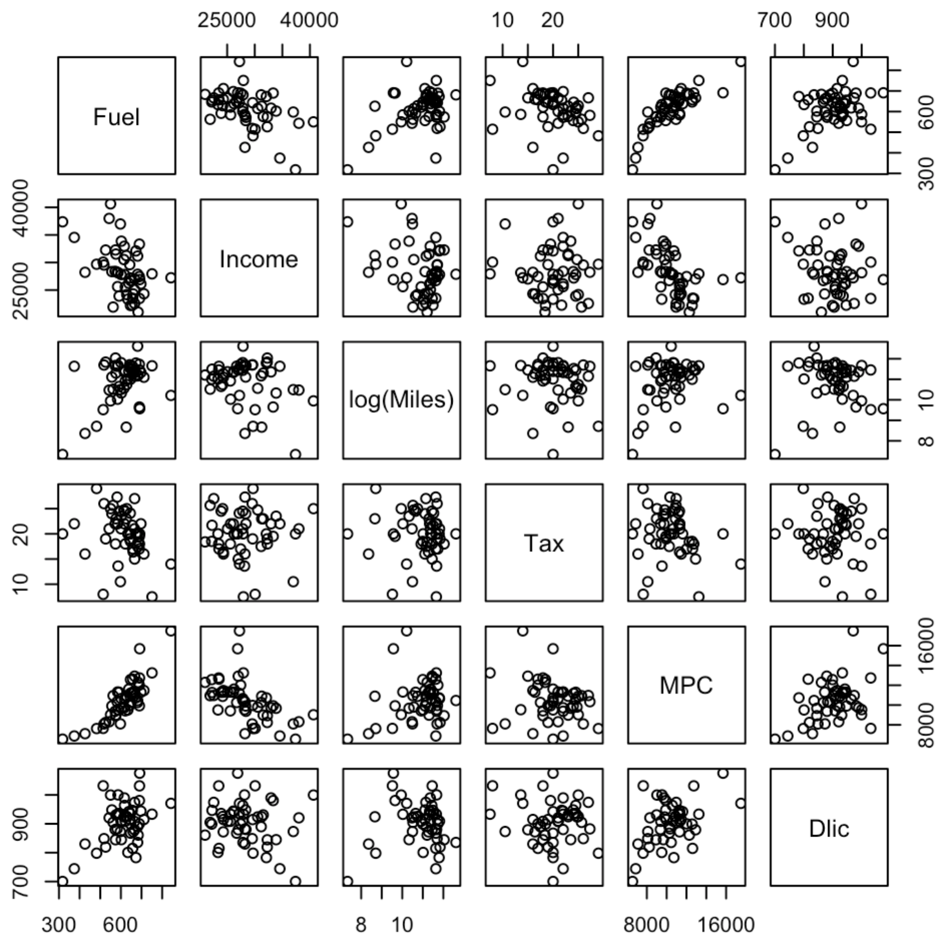
- Fitted: $E(\mathbf{Y}|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2$

Example: Fuel consumption

How does fuel consumption vary over the 50 states and the District of Columbia?

Variable	Description
Drivers	Number of licensed drivers in the state
FuelC	Gasoline sold for road use (thousands of gallons)
Income	Per person personal income (thousands of dollars)
Miles	Miles of Federal-aid highway miles in the state
Pop	Population age 16 and over
Tax	Gasoline state tax rate (cents per gallon)
MPC	Estimated miles driven per capita
Dlic	Number of licensed drivers in the state per 1,000 people age 16 and over ($1000 \times Drivers/Pop$)
Fuel	Fuel consumption (thousands of gallons) per 1,000 people age 16 and over ($1000 \times FuelC/Pop$)

```
pairs(Fuel ~ Income + log(Miles) + Tax + MPC + Dlic, data = fuel2001)
```



```r
mod1 <- lm(Fuel ~ Tax + Dlic + Income + log(Miles), data = fuel2001)
summary(mod1)
```

```
##
## Call:
## lm(formula = Fuel ~ Tax + Dlic + Income + log(Miles), data = fuel2001)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -163.145  -33.039    5.895   31.989  183.499
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 154.192845 194.906161   0.791 0.432938
## Tax          -4.227983   2.030121  -2.083 0.042873 *
## Dlic          0.471871   0.128513   3.672 0.000626 ***
## Income       -0.006135   0.002194  -2.797 0.007508 **
## log(Miles)   26.755176   9.337374   2.865 0.006259 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 64.89 on 46 degrees of freedom
## Multiple R-squared:  0.5105, Adjusted R-squared:  0.4679
## F-statistic: 11.99 on 4 and 46 DF,  p-value: 9.331e-07
```

Model checking: Linearity

Effects of a violation

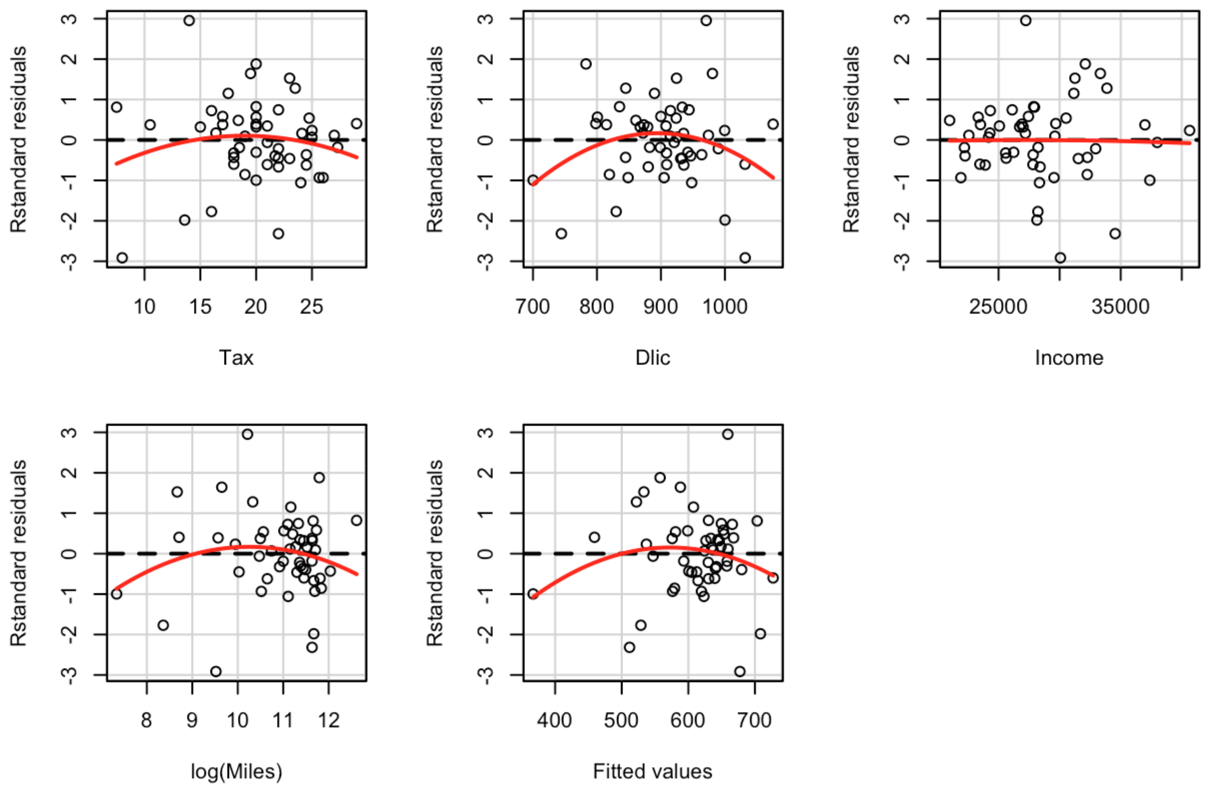
- coefficients and fitted values are biased
- inferences not valid

Diagnostics

- Residual plots
  - residuals vs. fitted values
  - residuals vs. each predictor
  - residuals vs. hypothetical new predictors
- Added variable plots
- Marginal model plots
- Test for curvature

Model checking: Linearity

```
residualPlots(mod1, layout = c(2, 3), type = "rstandard")
```



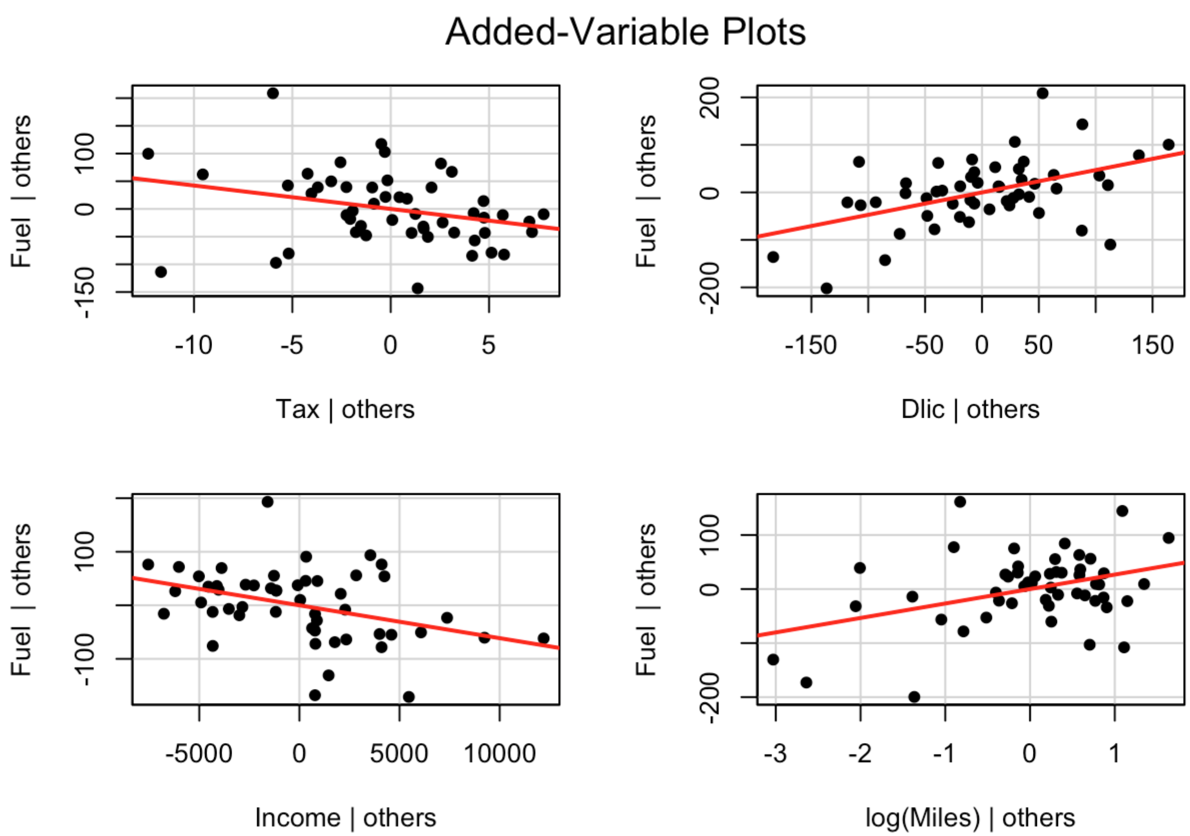
Model checking: Linearity

Added Variable Plots

- Display relationship of Y and X_j after adjusting for other variables
- Construction
  - $e_j =$ residuals regressing Y on all Xs except X_j
  - $u_j =$ residuals regressing X_j on all other Xs
  - Plot e_j vs. u_j

Added Variable Plots in R

```
avPlots(mod1, pch = 16)
```

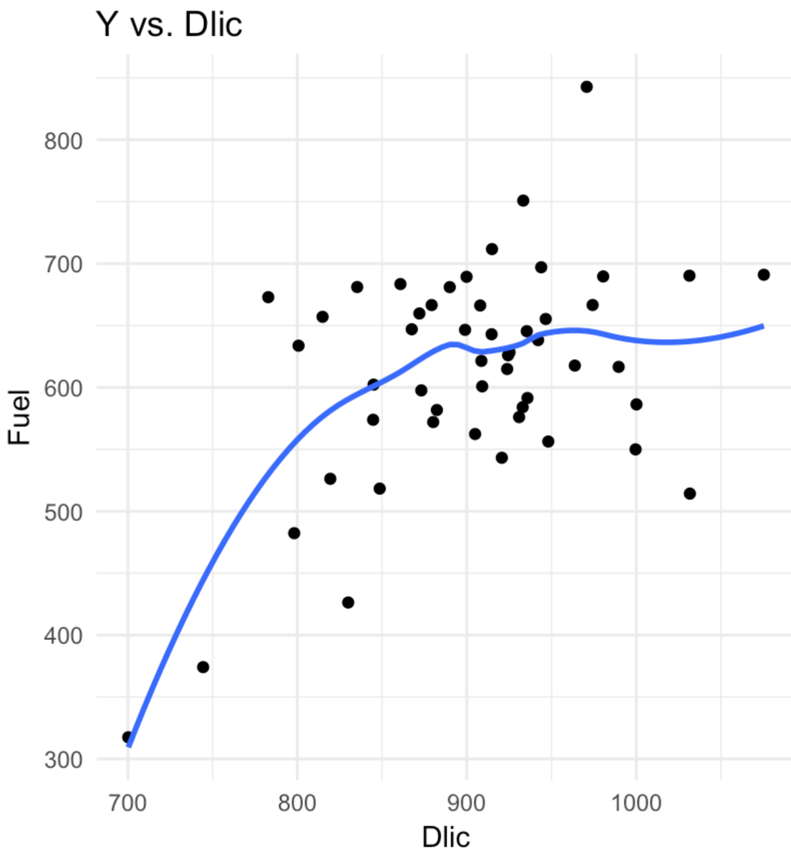


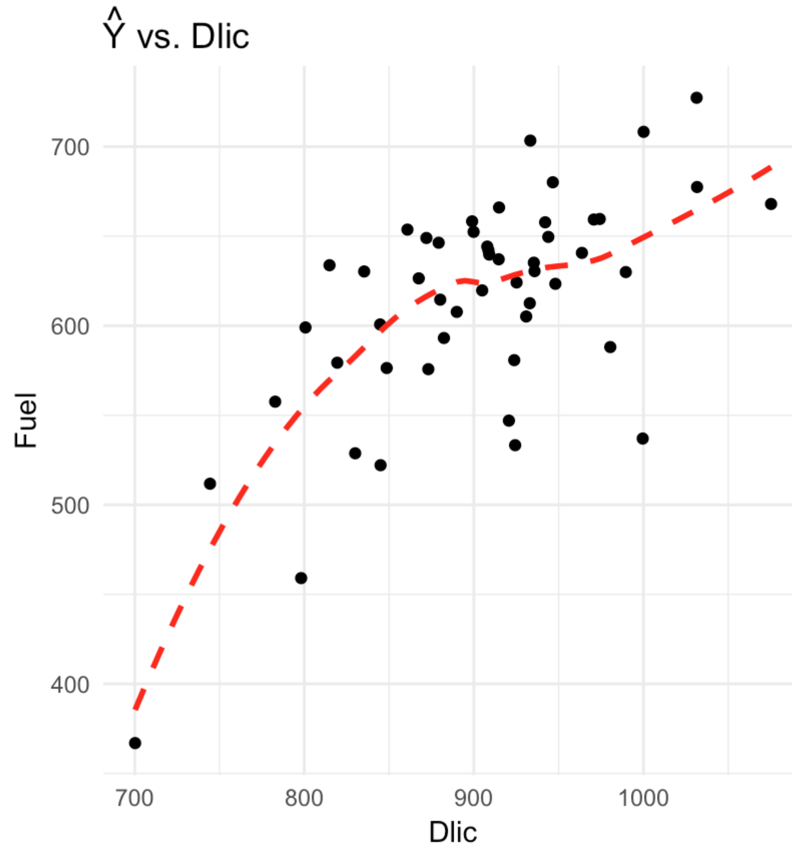
Model checking: Linearity

Marginal Model Plots

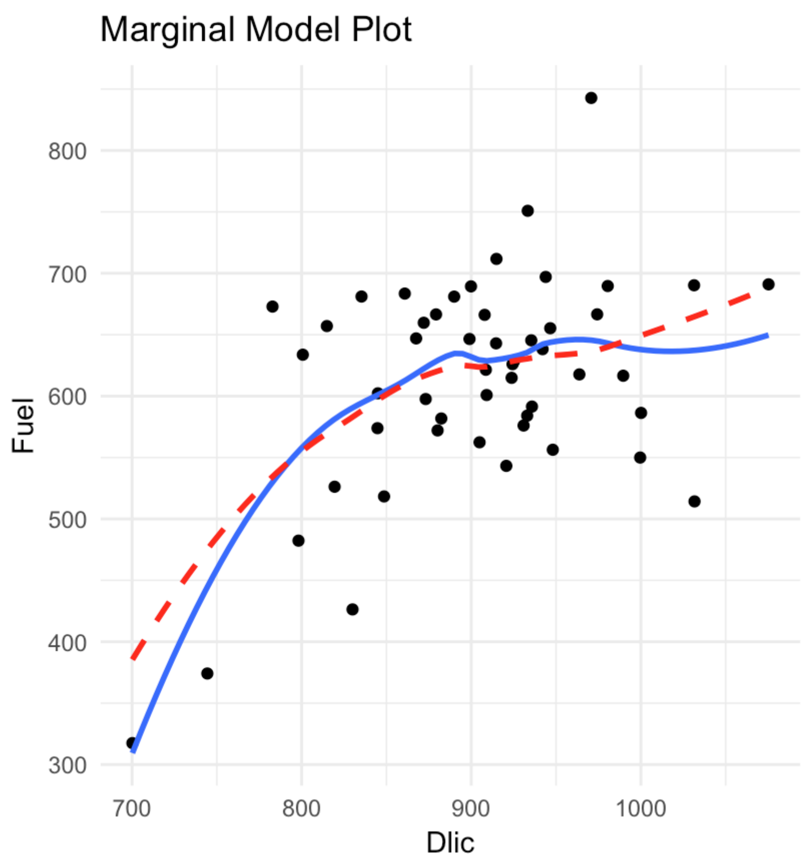
- Idea: check whether $E(Y|x_i)$ is well approximated by $E(\widehat{Y}|x_i)$
- Construction
  - Plot the response on the y-axis
  - Plot a predictor, or linear combination of the predictors (e.g. $\hat{y}_i$) on the x-axis
  - Add a nonparametric smoother to the plot
  - Use the fitted values from the model to plot the conditional mean (via nonparametric smoother)

Marginal Model Plots



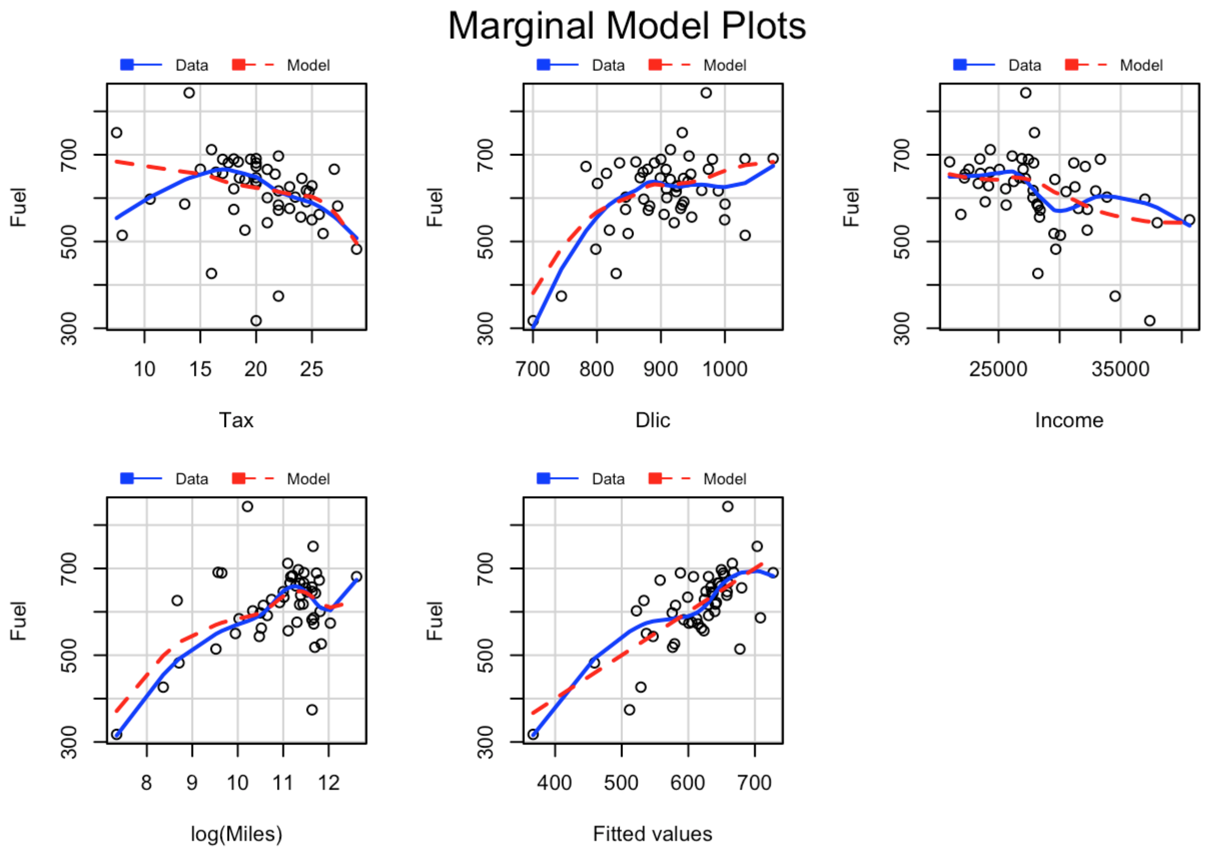


Marginal Model Plots



Marginal Model Plots in R

```
mmps(mod1, layout = c(2,3))
```



Marginal model vs. added variable plots

- **Marginal model plots**: are useful in checking to see that you're doing a good job of modeling the marginal relationship between a given predictor and the response.
- **Added variable plots**: assess how much variation in the response can be explained by a given predictor after the other predictors have already been taken into account (links to p-values).

Model checking: Linearity

Test for curvature

- If predictor U is in question, refit the regression with an added U^2 term.
- Look at t-test for the slope associated with U^2.

Test for curvature in R

```r
# I only specified plot = FALSE for the slides
residualPlots(mod1, plot = FALSE)
```

```
##            Test stat Pr(>|t|)
## Tax           -1.077    0.287
## Dlic          -1.922    0.061
## Income        -0.084    0.933
## log(Miles)    -1.347    0.185
## Tukey test    -1.446    0.148
```

Remedies for nonlinearity

- Nonlinear regression models (Grad school)
- Transformations of Y
- Transformations of X
- Polynomial regression

Model checking: Constant variance

Effects of a violation

- Coefficients are unbiased
- $se(\widehat{\beta}_i)$s are biased (often too small $\Longrightarrow$ CIs too narrow)

Diagnostics

- Residual plots
  - residuals vs. fitted values
  - residuals vs. each predictor
- Breusch-Pagan test

Model checking: Constant variance

Breusch-Pagan test

- H_0 : constant error variance
- H_A : error variance changes with the level of the response (fitted values), or with a linear combination of predictors

```
library(car) # if not loaded
ncvTest(mod1)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 0.03196885    Df = 1     p = 0.858096
```

```
ncvTest(mod1, ~ Tax + Dlic + Income + log(Miles))
```

```
## Non-constant Variance Score Test
## Variance formula: ~ Tax + Dlic + Income + log(Miles)
## Chisquare = 17.1249    Df = 4     p = 0.001827869
```

Remedies for nonconstant variance

- Transformations of Y
  - $\sqrt{Y}$ for counts
  - $\log$ of positive #s with large range
  - $\sin^{-1}(\sqrt{Y})$ or $\log\left(\frac{Y}{1-Y}\right)$ for proportions ($0 \leq Y \leq 1$)
- Weighted least squares (nice idea, see chapter 4)
- Use "sandwich estimator" for to obtain ses (inefficient)
- Use a generalized linear model

Model checking: Uncorrelated errors

Effects of a violation

- coefficients are unbiased
- $se(\widehat{\beta}_i)$s are biased, often smaller than necessary
- inferences not valid

Diagnostics

- Residual plots
    - residuals vs. time (or other factor inducing correlation)
    - residuals vs. lagged residuals
    - examine residuals in clusters (e.g., family, school)
- Think about data collection process

Model checking: Normal errors

Effects of a violation

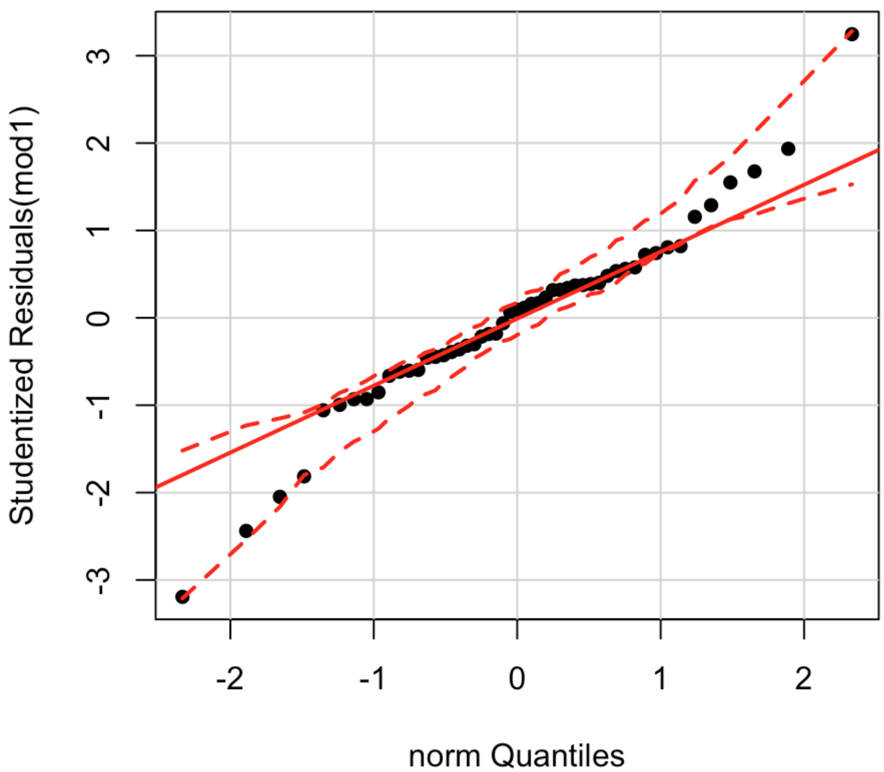
- problems with prediction
- inference for coefficients OK in large samples
- inference for coefficients problematic in small samples

Diagnostics

- normal Q-Q plot
- case diagnostics (for outliers)

Model checking: Normal errors

```
qqPlot(mod1, dist = "norm", pch = 16, line = "quartiles")
```



Remedies for non-normality

- Transformations
- Generalized linear models (Grad school or IS)

Detecting outliers and influential points

Standardized residuals

Recall that standardized residuals are calculated by dividing by their standard deviation

$$r_i = \frac{\widehat{e}_i}{s\sqrt{1 - h_{ii}}}$$

Observations with large standardized residuals can be considered outliers.

Rule of thumb:

- $|r_i| > 2$ for small data sets
- $|r_i| > 4$ for large data sets

Case diagnostics: Leverage

Definition

Pull off the diagonal elements of the hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Cutoff

- $2(k + 1)/n$
- $3(k + 1)/n$
- Also a good idea to examine a histogram

Case diagnostics: DFFITS

Measures the effect of the i^{th} case on the fitted value for Y_i

$$DFFITS_i = \frac{\widehat{Y}_i - \widehat{Y}_{i(i)}}{\sqrt{\text{MSE}_{(i)} h_{ii}}}$$

where $MSE_{(i)} = RSS_{(i)}/(n - p - 1)$

Cutoff

- $|DFFITS_i| > 1$ considered large in small or medium samples
- $|DFFITS_i| > 2\sqrt{\frac{p+1}{n}}$ considered large in big samples
- Judge by relative standing

Case diagnostics: Cook's D

Measures effect an observation has on <u>all</u> of the fitted values

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{Y}_j - \widehat{Y}_{j(i)} \right)^2}{(p+1)\text{MSE}} = \frac{r_i^2}{p+1} \frac{h_{ii}}{1-h_{ii}}$$

Cutoff

- $D_i > F_{p+1,n-p-1,0.5}$ is of substantial concern
- Also can judge relative standing of D_i

Case diagnostics: DFBETAS

Measures the effect of an observation on a single coefficient

$$DFBETAS_{k,i} = \frac{\widehat{\beta}_k - \widehat{\beta}_{k(i)}}{SE_{\beta_{(i)}}}$$

Cutoff

- $DFBETA_i > 1$ in small or medium samples
- $DFBETA_i > 2/\sqrt{n}$ in large samples

Case diagnostics in R

```
case.infl <- influence.measures(mod1)

# print which observations "are" influential
which(apply(case.infl$is.inf, 1, any))
```

```
##  2  7  9 31 33 40 51
##  2  7  9 31 33 40 51
```

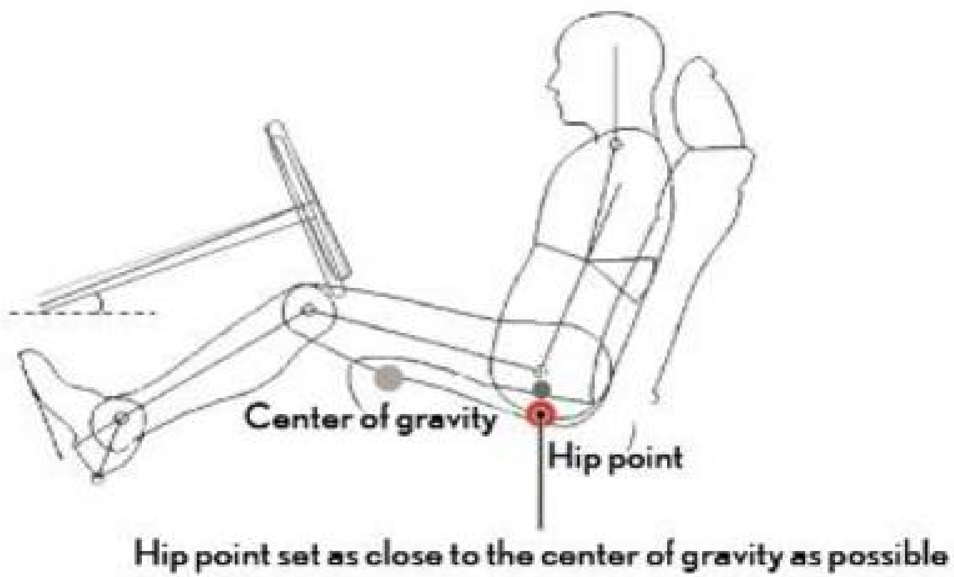
```
# We can also call the measure individually
hatvalues(mod1)
cooks.distance(mod1)
dffits(mod1)
dfbetas(mod1)
```

Detecting multicollinearity

Example: Car seat position

Research question:

- Car designers would find it helpful to know where different drivers will position the seat depending on their size and age.
- Position can be measured by hip center



Example: Car seat position

Researchers at the HuMoSim laboratory at the University of Michigan collected data on 38 drivers:

Variable	Description
`Age`	Age (years)
`Weight`	Weight (lbs)
`HtShoes`	Height in shoes (cm)
`Ht`	Height in bare feet (cm)
`Seated`	Seated height (cm)
`Arm`	Lower arm length (cm)
`Thigh`	Thigh length (cm)
`Leg`	Lower leg length (cm)
`hipcenter`	horizontal distance of the midpoint of the hips from a fixed location in the car (mm)

Example: Car seat position

```
##   Age Weight HtShoes    Ht Seated  Arm Thigh  Leg hipcenter
## 1  46    180   187.2 184.9   95.2 36.1  45.3 41.3  -206.300
## 2  31    175   167.5 165.5   83.8 32.9  36.5 35.9  -178.210
## 3  23    100   153.6 152.2   82.9 26.0  36.6 31.0   -71.673
## 4  19    185   190.3 187.4   97.3 37.4  44.1 41.0  -257.720
## 5  23    159   178.0 174.1   93.9 29.5  40.1 36.9  -173.230
## 6  47    170   178.7 177.0   92.4 36.0  43.2 37.4  -185.150
```

Example: Car seat position

```
seatpos_mod <- lm(hipcenter ~ ., data = seatpos)
# the dot in the formula notation means 'all other variables'
broom::tidy(seatpos_mod)
```

```
##          term     estimate   std.error   statistic    p.value
## 1 (Intercept) 436.43212823 166.5716187  2.62008697 0.01384361
## 2         Age   0.77571620   0.5703288  1.36012113 0.18427175
## 3      Weight   0.02631308   0.3309704  0.07950283 0.93717877
## 4     HtShoes  -2.69240774   9.7530351 -0.27605845 0.78446097
## 5          Ht   0.60134458  10.1298739  0.05936348 0.95306980
## 6      Seated   0.53375170   3.7618942  0.14188376 0.88815293
## 7         Arm  -1.32806864   3.9001969 -0.34051323 0.73592450
## 8       Thigh  -1.14311888   2.6600237 -0.42974011 0.67056106
## 9         Leg  -6.43904627   4.7138601 -1.36598163 0.18244531
```

```
broom::glance(seatpos_mod)
```

```
##   r.squared adj.r.squared    sigma statistic      p.value df    logLik
## 1 0.6865535     0.6000855 37.72029  7.939971 1.305773e-05  9 -186.7317
##        AIC      BIC deviance df.residual
## 1 393.4634 409.8392 41261.78          29
```

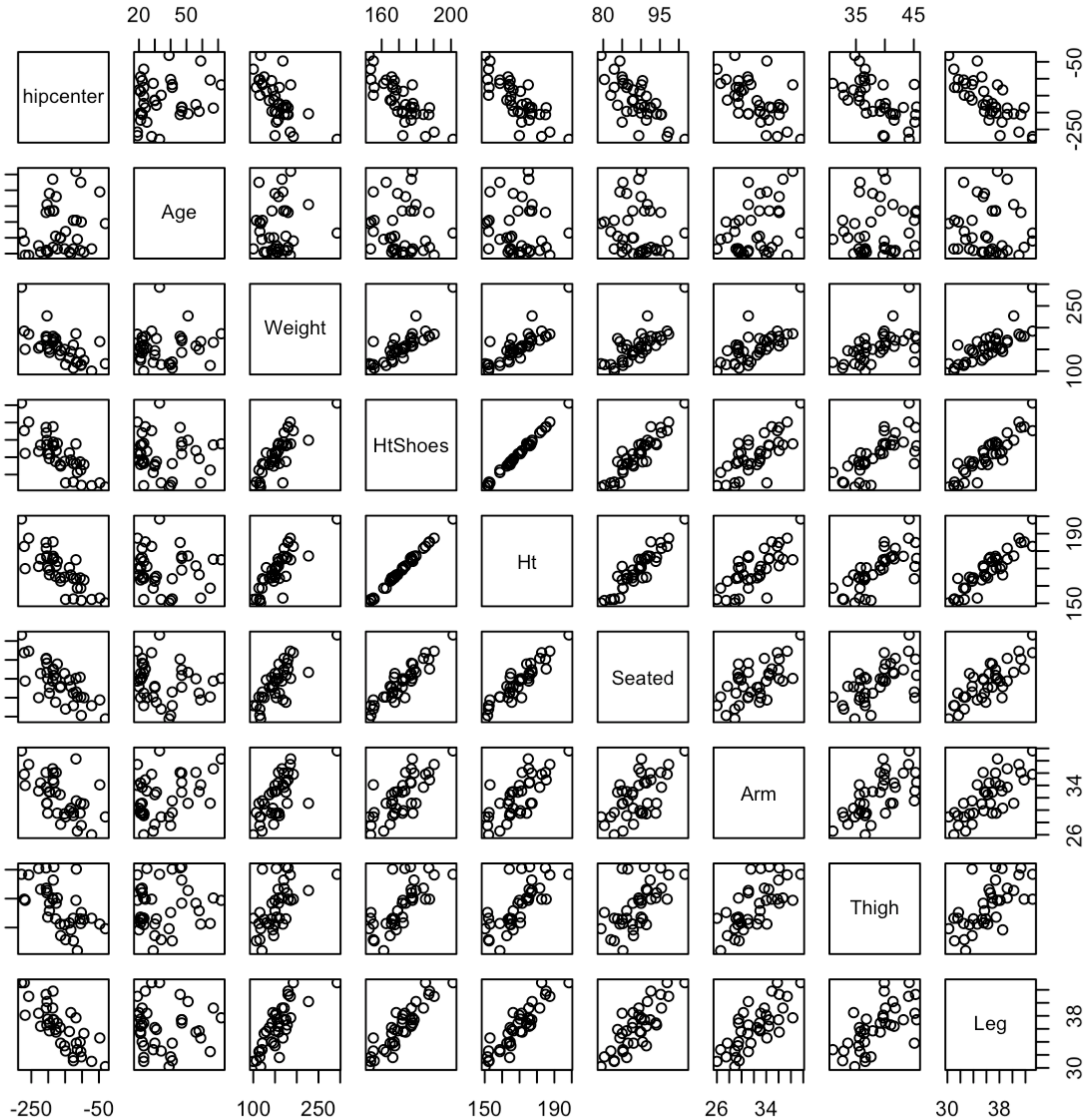
Multicollinearity

- Definition: situation where 1+ predictors are "nearly" linearly related to the others.

- This is not a model violation, but we treat it like one because…
  - $se(\widehat{\beta}_i)$s are too big
  - it's difficult to interpret individual $se(\widehat{\beta}_i)$s
  - $se(\widehat{\beta}_i)$s change a lot with minor changes to the model/data (e.g., if we remove a case or a variable)

Diagnosing Multicollinearity

- Examine scatterplot matrix
- Examine pairwise correlations between predictors
- Look for $\widehat{\beta}_i$s with unusual signs
- Notice great sensitivity
- Calculate the **Variance Inflation Factor (VIF)**



hipcenter
Age
Weight
HtShoes
Ht
Seated
Arm
Thigh
Leg

```
cor(seatpos)
```

```
##                Age Weight HtShoes    Ht Seated   Arm  Thigh    Leg hipcenter
## Age          1.000  0.081  -0.079 -0.09  -0.17  0.36  0.091 -0.042      0.21
## Weight       0.081  1.000   0.828  0.83   0.78  0.70  0.573  0.784     -0.64
## HtShoes     -0.079  0.828   1.000  1.00   0.93  0.75  0.725  0.908     -0.80
## Ht          -0.090  0.829   0.998  1.00   0.93  0.75  0.735  0.910     -0.80
## Seated      -0.170  0.776   0.930  0.93   1.00  0.63  0.607  0.812     -0.73
## Arm          0.360  0.698   0.752  0.75   0.63  1.00  0.671  0.754     -0.59
## Thigh        0.091  0.573   0.725  0.73   0.61  0.67  1.000  0.650     -0.59
## Leg         -0.042  0.784   0.908  0.91   0.81  0.75  0.650  1.000     -0.79
## hipcenter    0.205 -0.640  -0.797 -0.80  -0.73 -0.59 -0.591 -0.787      1.00
```

Variance Inflation Factor (VIF)

The variance of a given slope can be written

$$Var(\widehat{\beta}_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma^2}{(n-1)s_{x_j}^2}$$

The first term is the VIF: $\frac{1}{1 - R_j^2}$

- $\text{VIF}_j > 5$ suspicions begin ($R_i^2 > .8$)
- $\text{VIF}_j > 10$ indicates a problem ($R_i^2 > .9$)
- $\text{VIF}_j > 100$ indicates a big problem ($R_i^2 > .99$)

Diagnosing Multicollinearity

```
vif(seatpos_mod)
```

```
##        Age     Weight    HtShoes         Ht     Seated        Arm
##   1.997931   3.647030 307.429378 333.137832   8.951054   4.496368
##      Thigh        Leg
##   2.762886   6.694291
```

Remedies for collinearity

- Use model only for prediction
- Drop highly correlated variables (USE CAUTION!)
- Create composite variables
- Find new cases that "break" the observed correlation (i.e., have a different pattern)

Simplifying the model

```
seatpos_mod2 <- lm(hipcenter ~ Age + Weight + Ht, data = seatpos)
summary(seatpos_mod2)
```

```
##
## Call:
## lm(formula = hipcenter ~ Age + Weight + Ht, data = seatpos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -91.526 -23.005   2.164  24.950  53.982
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 528.297729 135.312947   3.904 0.000426 ***
## Age           0.519504   0.408039   1.273 0.211593
## Weight        0.004271   0.311720   0.014 0.989149
## Ht           -4.211905   0.999056  -4.216 0.000174 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.49 on 34 degrees of freedom
## Multiple R-squared:  0.6562, Adjusted R-squared:  0.6258
## F-statistic: 21.63 on 3 and 34 DF,  p-value: 5.125e-08
```

Simplifying the model

```
vif(seatpos_mod2)
```

```
##      Age   Weight       Ht
## 1.093018 3.457681 3.463303
```

Diagnostics summary

Diagnostics summary

Before drawing any conclusions from a regression model, we must be confident it is a valid way to model the data. Our model assumes:

- **Linearity:** the conditional mean of the response is a linear function of the predictors.
- The errors have **constant variance** and are **uncorrelated.**
- The errors are **normally distributed with mean zero**.

It's also sensible to build a model with:

- No highly influential points.
- Low multicollinearity.