LINEAR REGRESSION

ЛИНЕЙНАЯ РЕГРЕССИЯ

LINEAR REGRESSION

REGRESSION AND OTHER MODELS

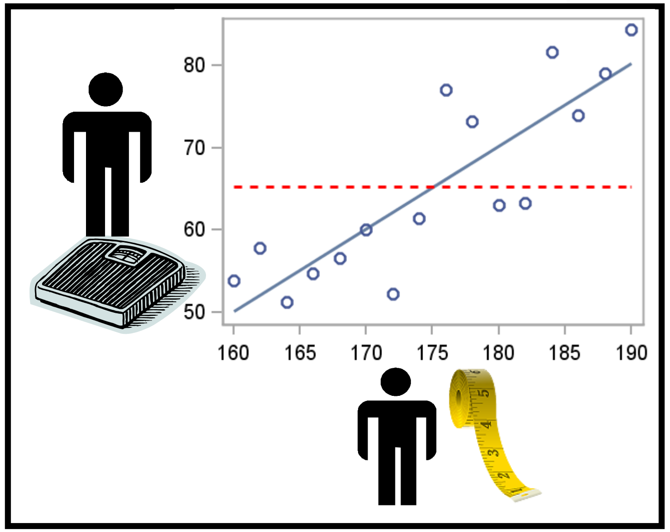
Type of Predictors / Type of Response	Categorical категориальный	Continuous непрерывный	Continuous and Categorical
Continuous непрерывный	Analysis of Variance (ANOVA)	**Ordinary Least Squares (OLS) Regression**	Analysis of Covariance (ANCOVA)
Categorical категориальный	Contingency Table Analysis or Logistic Regression	Logistic Regression	Logistic Regression

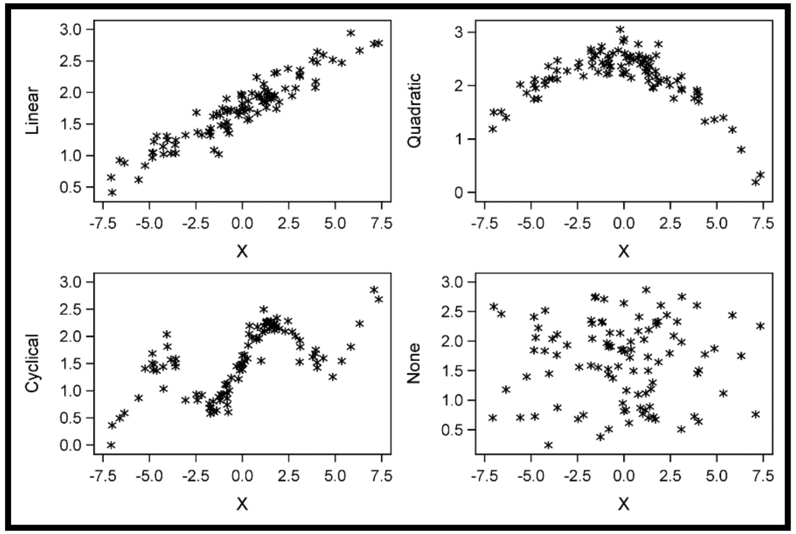
LINEAR REGRESSION

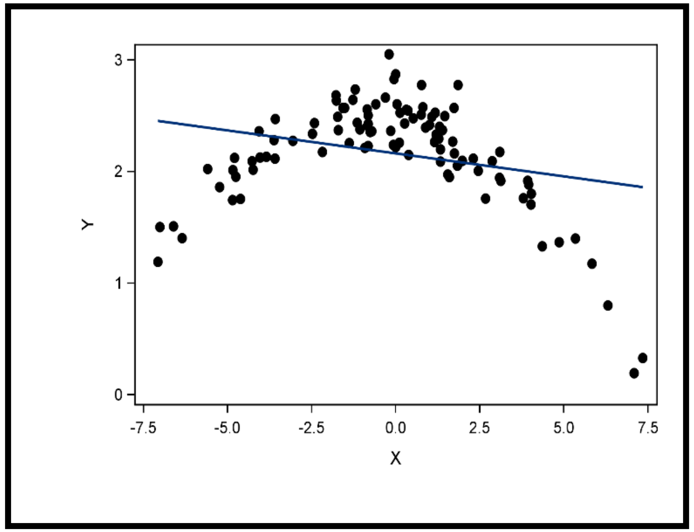
...REGRESSIONS...

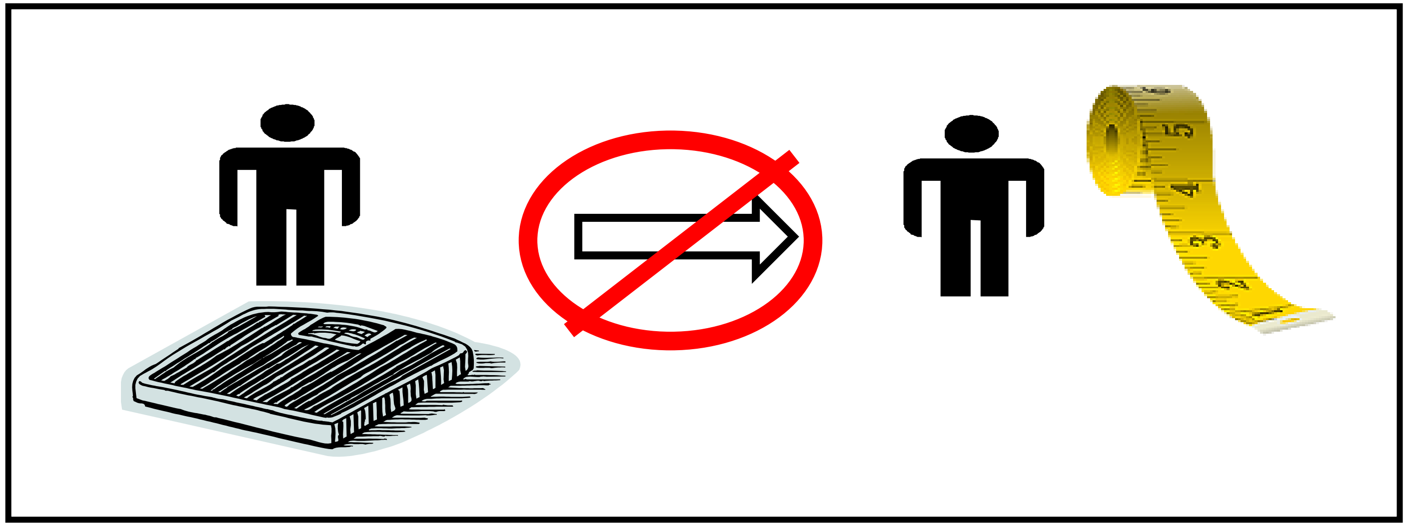
- **linear** / non-linear
- logistic
- OLS
- PLS
- LAR
- RIDGE
- LASSO
- LOESS
- ROBUST
- QUANTILE
- ...

RELATIONSHIP | HEIGHT-WEIGHT – CORRELATION???









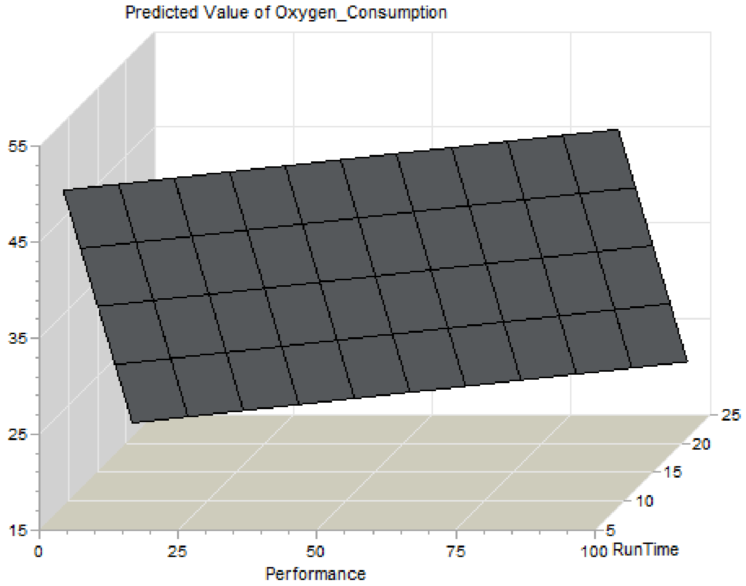
MULTIPLE LINEAR REGRESSION

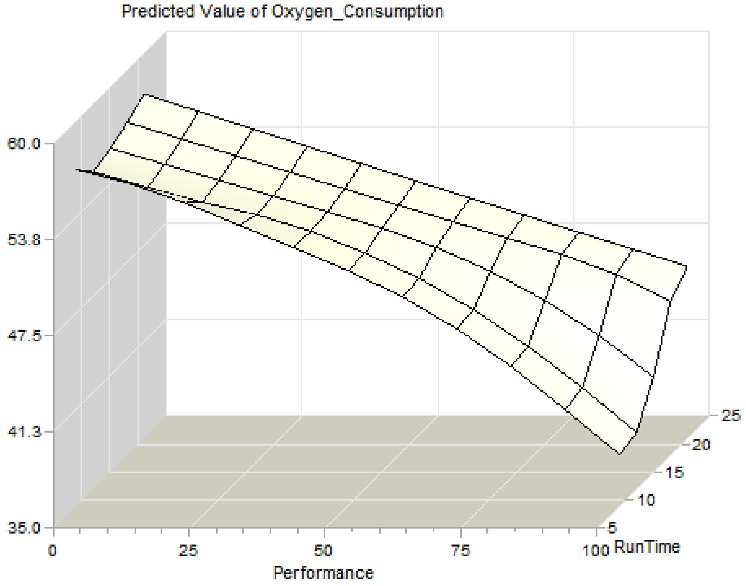
МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

MULTIPLE LINEAR REGRESSION

MODEL

- Обычно, вы моделируете зависимую переменную Y, линейную функцию от *k* независимых переменных $X_1 \ldots X_k$:
- $Y=\beta_0+\beta_1 X_1+\ldots+\beta_k X_k+\varepsilon$





$Y=\beta_0+\beta_1 X_1+\beta_2 X_2+\varepsilon$

Linear?

$Y=\beta_0+\beta_1 X_1+\beta_2 X_1^2+\beta_3 X_2+\beta_4 X_2^2+\varepsilon$

Nonlinear?

MULTIPLE LINEAR REGRESSION

SAS/STAT SYNTAX & EXAMPLE DATA

```
proc reg data=sasuser.fitness;
    MODEL Oxygen_Consumption = RunTime
                               Age
                               Weight
                               Run_Pulse
                               Rest_Pulse
                               Maximum_Pulse
                               Performance;
run; quit;
```

MULTIPLE LINEAR REGRESSION

APPLICATIONS: ПРЕДСКАЗАНИЕ VS. ИССЛЕДОВАНИЕ

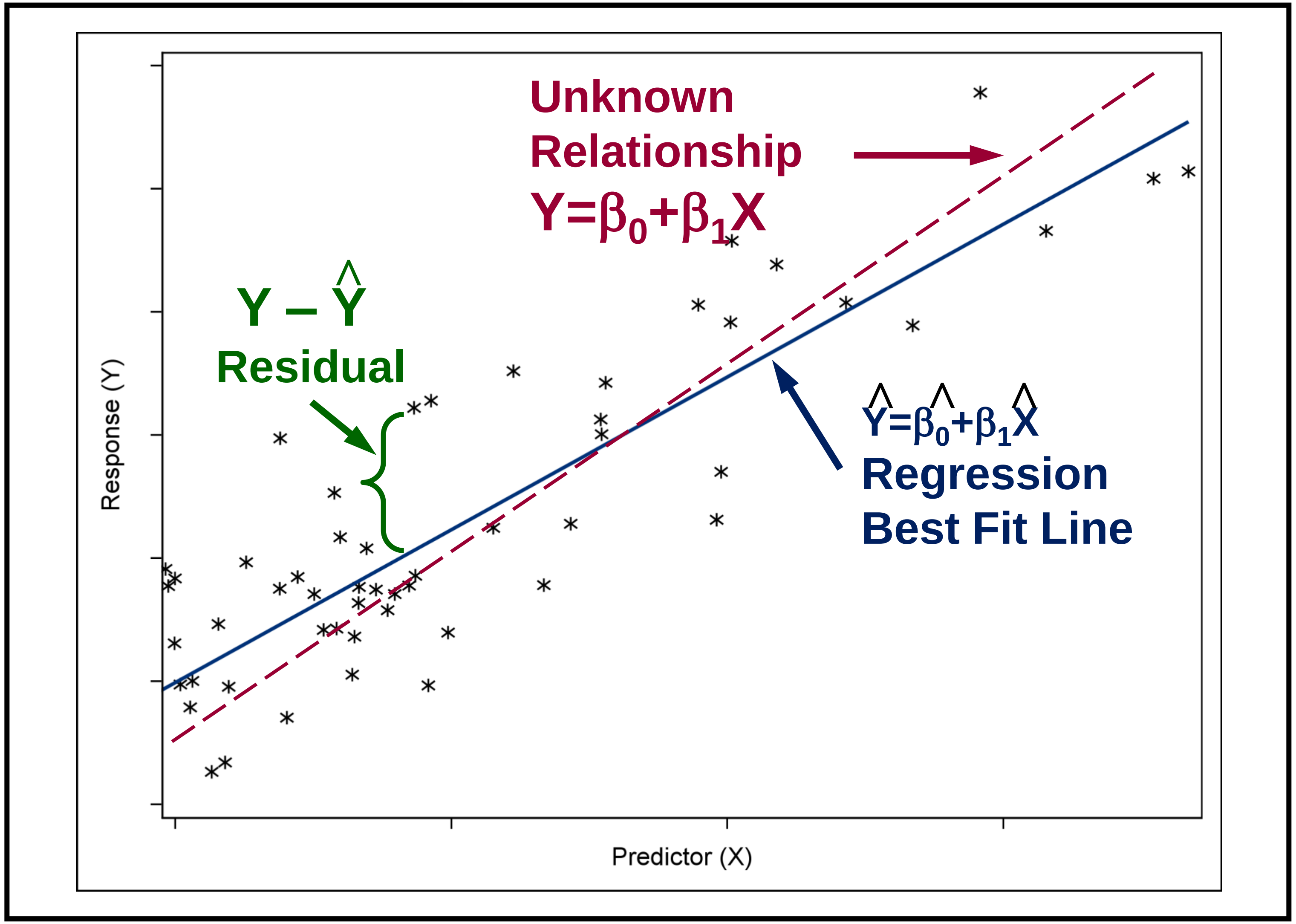
- Предикторы, их знаки и статистическая значимость представляют ***вторичный интерес***.
- **Фокусируемся на построении модели, лучшей с точки зрения предсказания будущих значений** Y, т.е. более точной модели.

$$\underline{\hat{Y}} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k$$

- **Фокусируемся на понимании взаимосвязи** между целевой (зависимой) переменной и предикторами (независимыми) переменными.
- Поэтому, **важна статистическая значимость предикторов**, а также **значения и знаки коэффициентов** в модели.

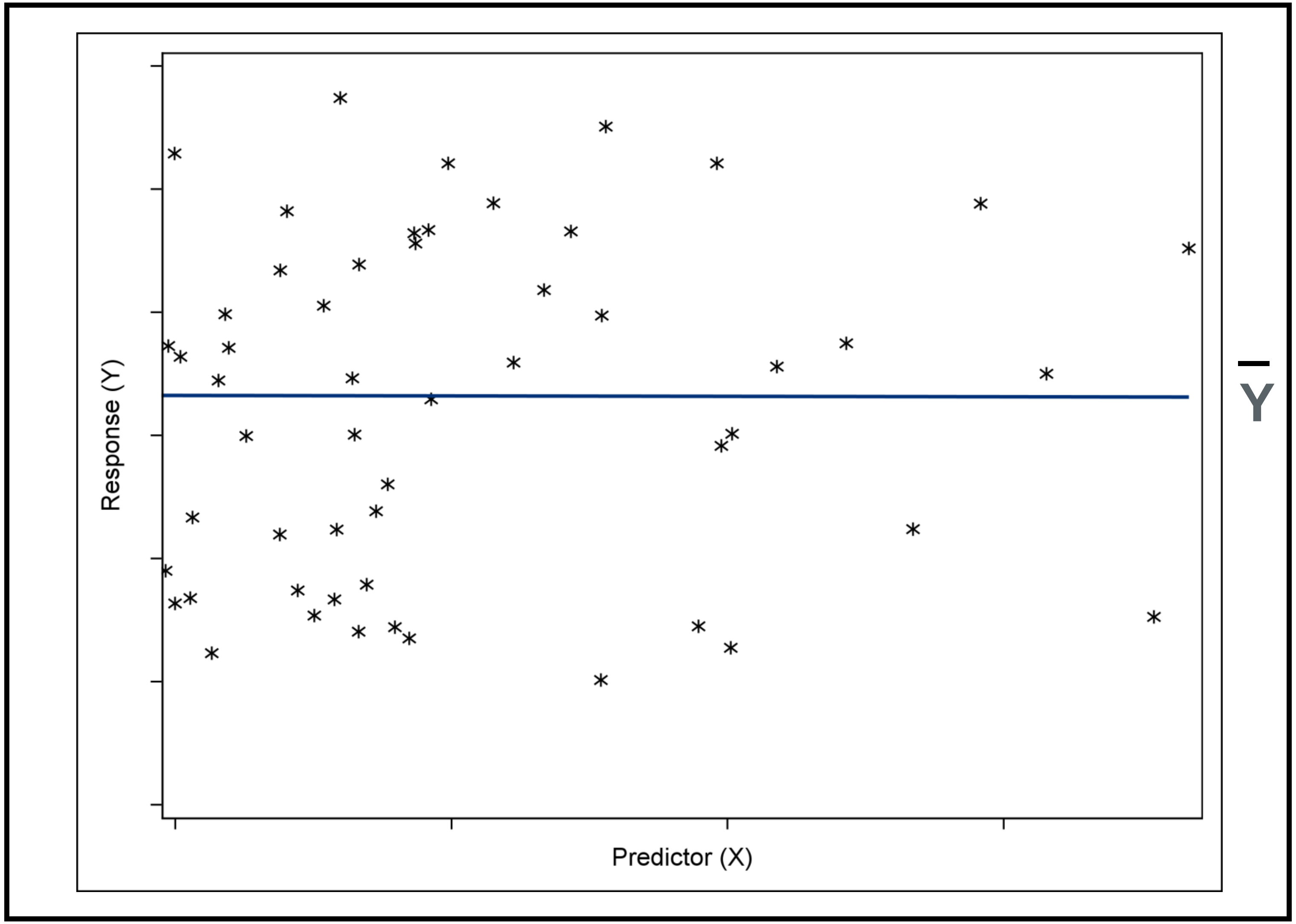
$$\hat{Y} = \underline{\hat{\beta}_0} + \underline{\hat{\beta}_1} X_1 + \ldots + \underline{\hat{\beta}_k} X_k$$

SIMPLE LINEAR REGRESSION MODEL



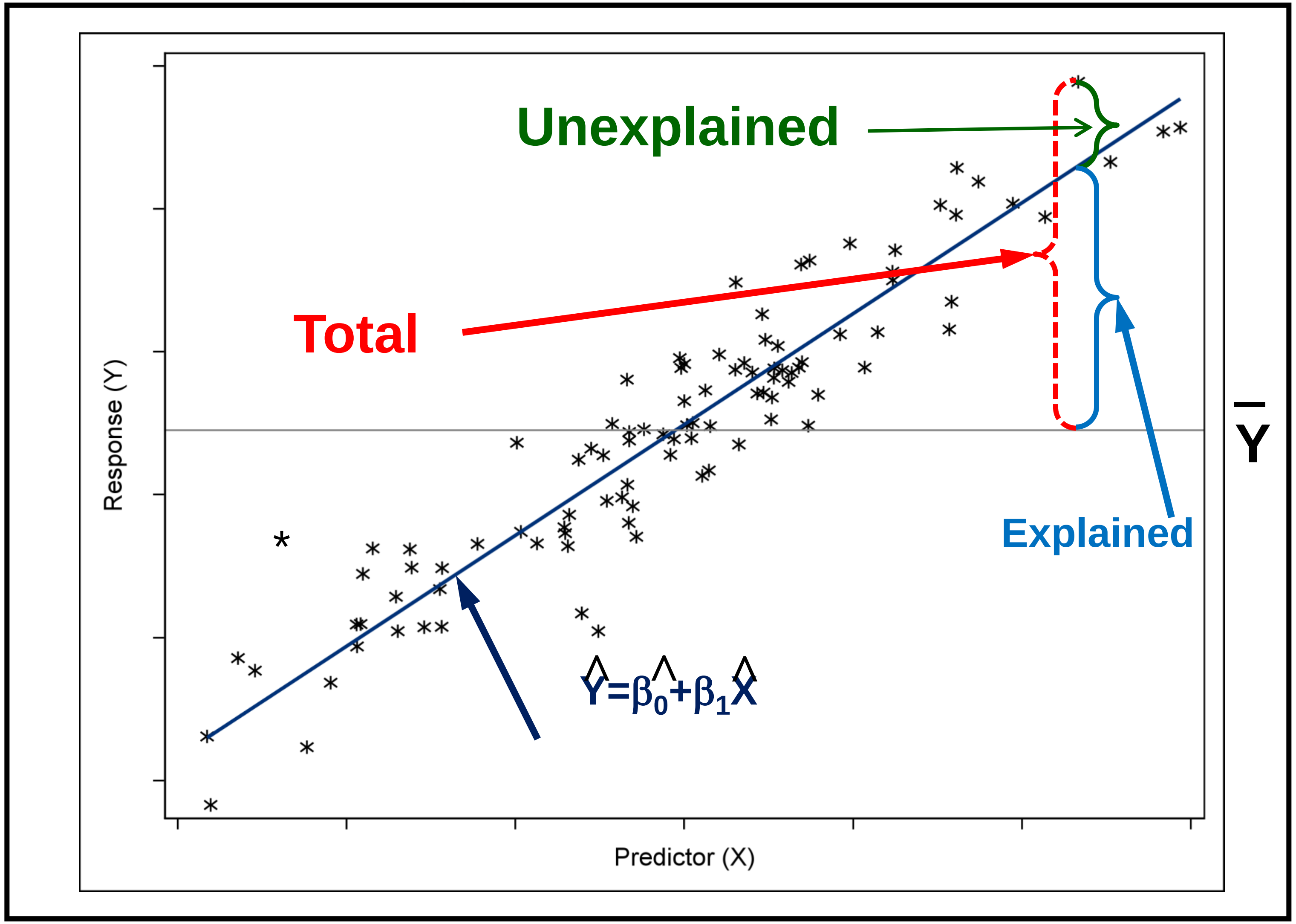
SIMPLE LINEAR REGRESSION

THE BASELINE MODEL – БАЗОВАЯ МОДЕЛЬ



SIMPLE LINEAR REGRESSION

VARIABILITY



Общая, объясненная и необъясненная вариативность

HYPOTHESES

Null Hypothesis:

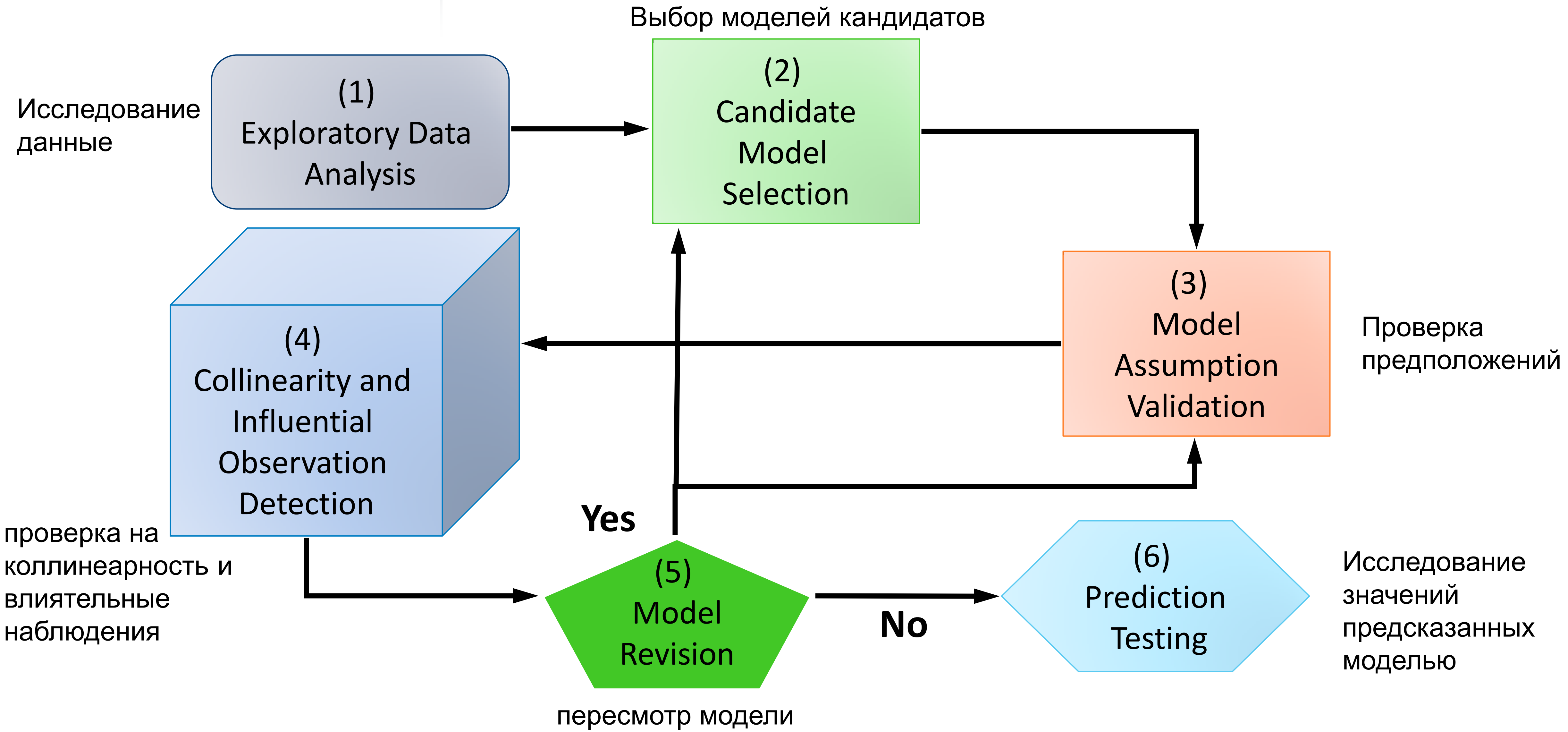
- The regression model does ***not*** fit the data better than the baseline model.
- $\beta_1=\beta_2=\ldots=\beta_k=0$ – F-statistic
  - Also $\beta_i=0$ for each predictor – t-statistic

Alternative Hypothesis:

- The regression model does fit the data better than the baseline model.
- Not all β_is equal zero.

MULTIPLE LINEAR REGRESSION

MODEL DEVELOPMENT PROCESS



(2) CANDIDATE MODEL SELECTION

MULTIPLE LINEAR REGRESSION

MODEL SELECTION

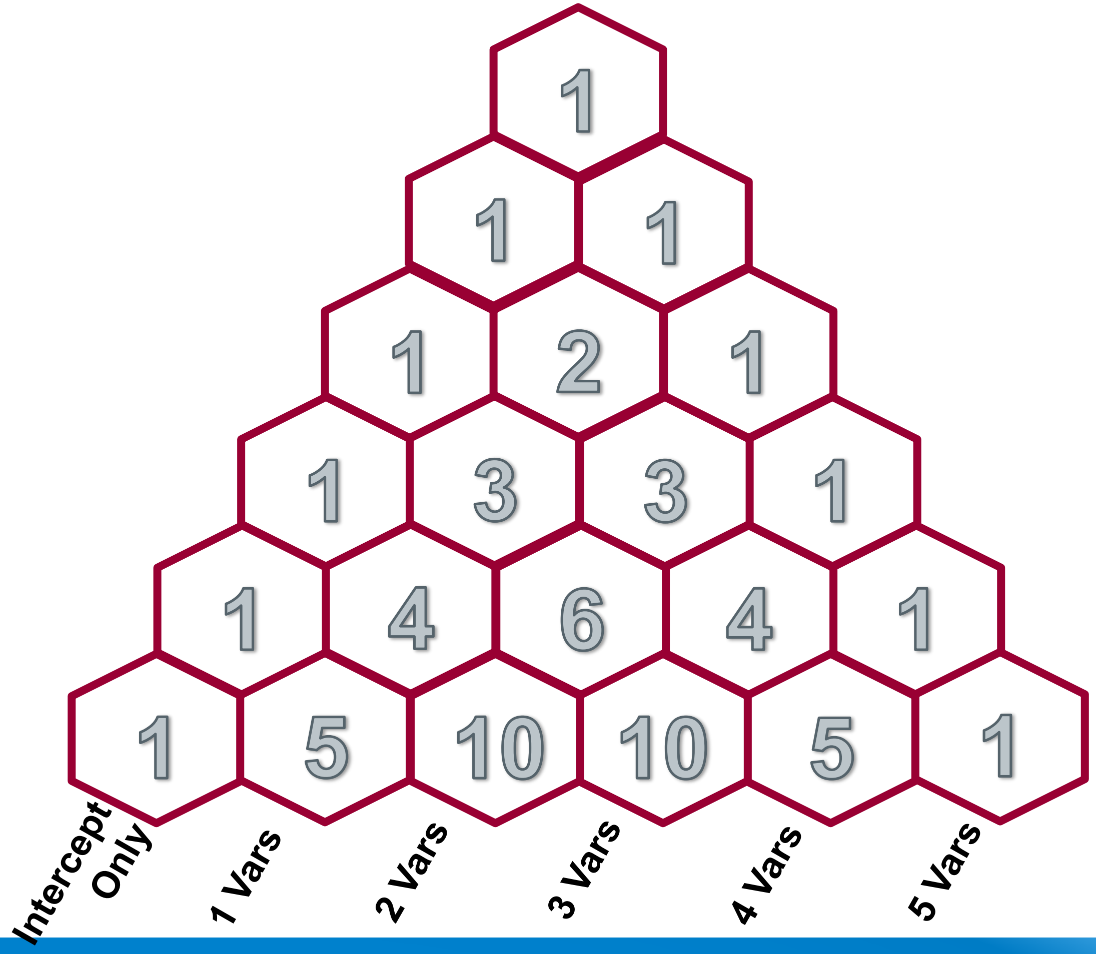
MODEL SELECTION OPTIONS

- The `SELECTION=` option in the MODEL statement of PROC REG supports these model selection techniques:
  - **Stepwise selection methods**
    - STEPWISE, FORWARD, or BACKWARD
  - **All-possible regressions ranked using**
    - RSQUARE, ADJRSQ, or CP
  - **MINR, MAXR** *[home work]*
  - SELECTION=NONE is the default.

CANDIDATE MODEL SELECTION

ALL-POSSIBLE REGRESSIONS

Variables in Full Model (k)	Total Number of Subset Models (2^k)
0	1
1	2
2	4
3	8
4	16
5	32



CANDIDATE MODEL SELECTION

ALL-POSSIBLE REGRESSIONS

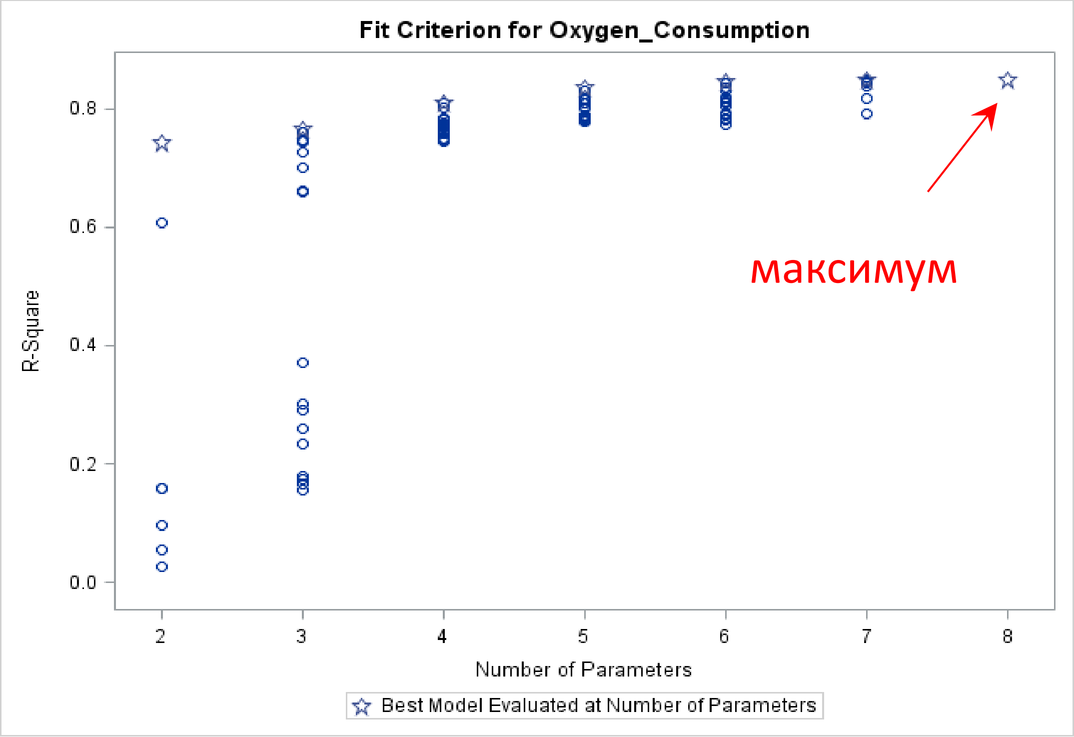
```
ods graphics / imagemap=on;

proc reg data=sasuser.fitness
         plots(only)=(rsquare adjrsq cp);
    ALL_REG: model oxygen_consumption
                    = Performance RunTime Age Weight
                      Run_Pulse Rest_Pulse Maximum_Pulse
            / selection=rsquare
                    adjrsq cp best=10;
    title 'Best Models Using All-Regression Option';
run;
quit;
```

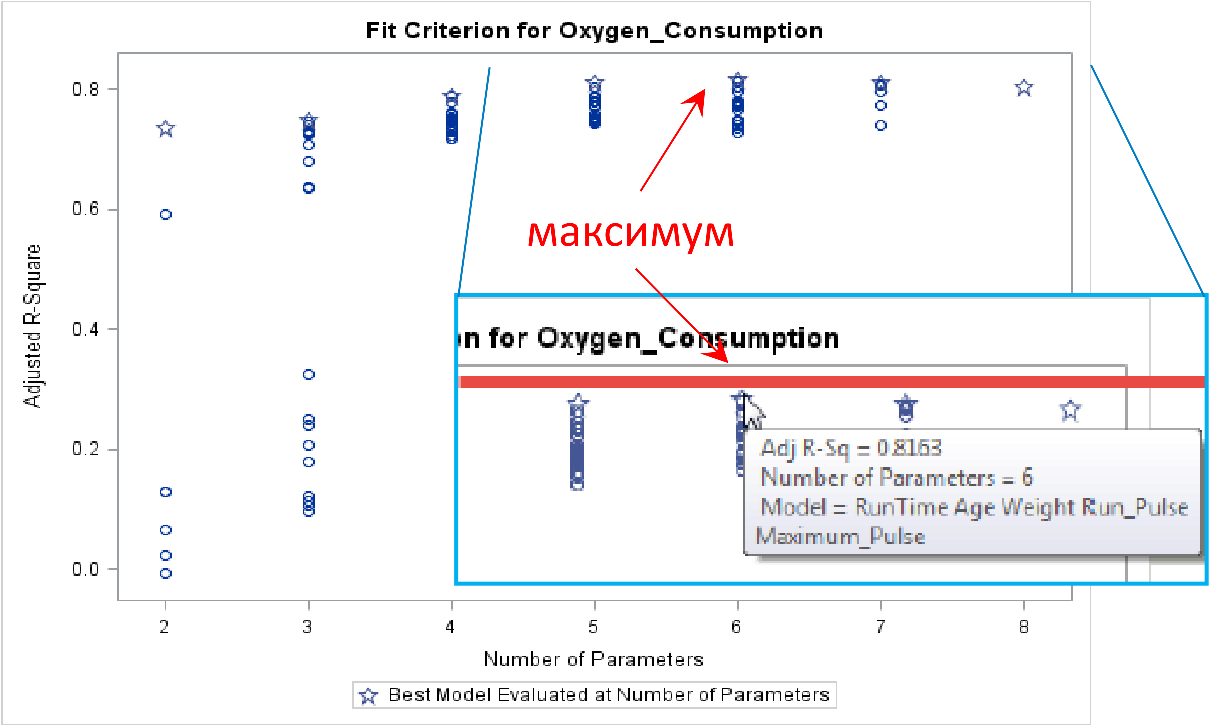
CANDIDATE MODEL SELECTION

ALL-POSSIBLE REGRESSIONS: RANK

$$R^2 = 1 - \frac{SS_E}{SS_T} = \frac{SS_M}{SS_T}$$

$$R^2_{ADJ} = 1 - \frac{(n-i)(1-R^2)}{n-p}$$





MODEL SELECTION | MALLOWS' C_P

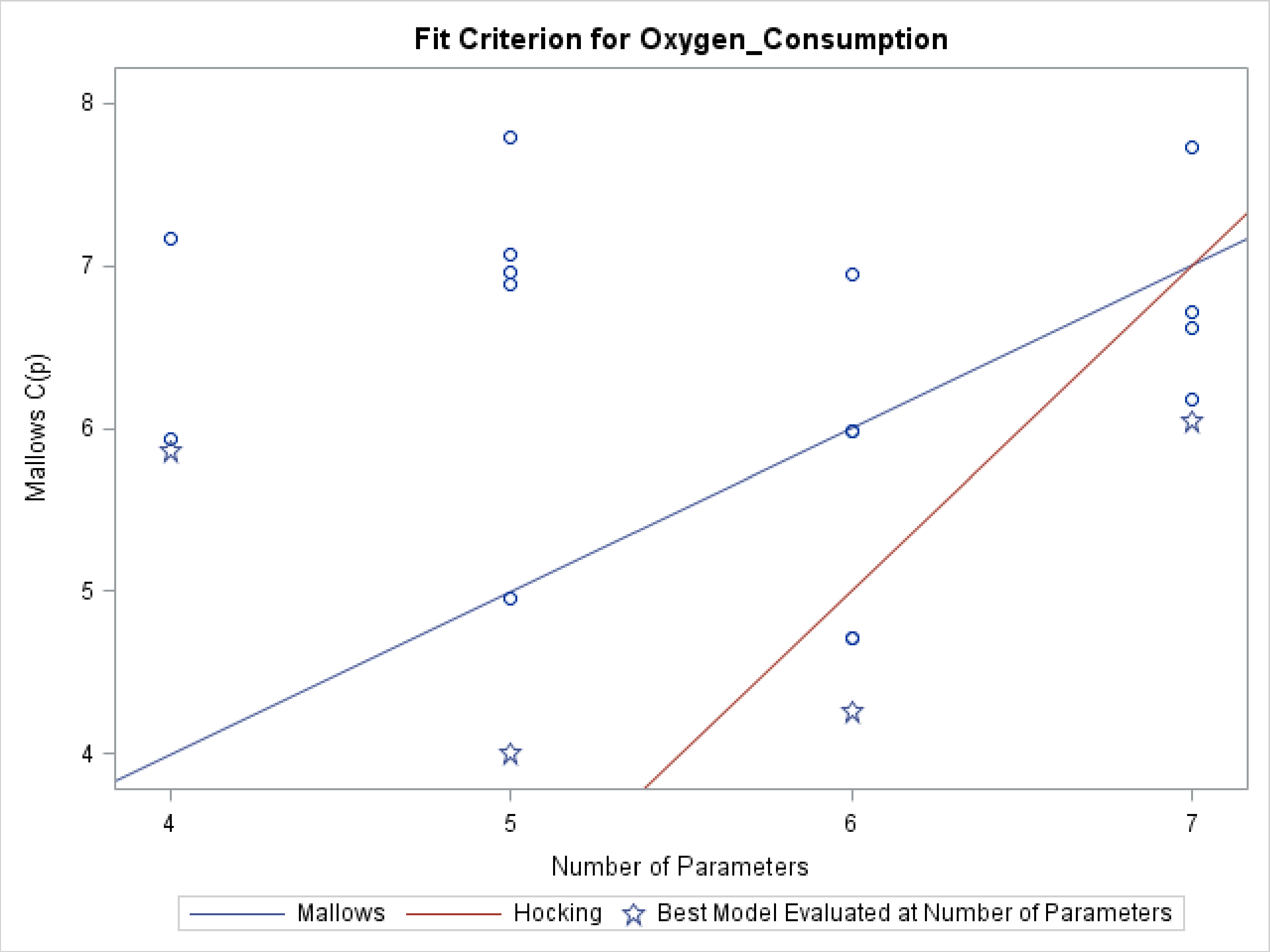
- Look for models with *max* p : $C_p \leq p$, *p* = number of parameters + intercept.

$$C_p = p + \frac{(MSE_p - MSE_{full})(n-p)}{MSE_{full}}$$

HOCKING'S CRITERION VERSUS MALLOWS' C_P

- Hocking (1976) suggests selecting a model based on the following:
  - $C_p \leq p$ for prediction
  - $C_p \leq 2p - p_{full} + 1$ for parameter estimation

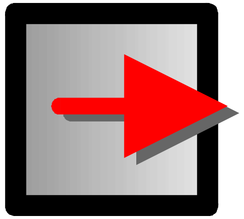
MODEL SELECTION | MALLOWS' C_P



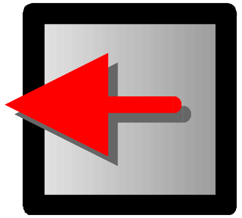
MODEL SELECTION ALL-POSSIBLE REGRESSIONS RANKED USING

Model Index	Number in Model	C(p)	R-Square	Adjusted R-Square	Variables in Model
1	4	4.0004	0.8355	0.8102	RunTime **Age** Run_Pulse **Maximum_Pulse**
2	5	4.2598	0.8469	0.8163	RunTime Age Weight Run_Pulse Maximum_Pulse
3	5	4.7158	0.8439	0.8127	Performance RunTime Weight Run_Pulse Maximum_Pulse
4	5	4.7168	0.8439	0.8127	Performance RunTime Age Run_Pulse Maximum_Pulse
5	4	4.9567	0.8292	0.8029	Performance RunTime Run_Pulse Maximum_Pulse
6	3	5.8570	0.8101	0.7890	RunTime Run_Pulse Maximum_Pulse
7	3	5.9367	0.8096	0.7884	RunTime Age Run_Pulse
8	5	5.9783	0.8356	0.8027	RunTime Age Run_Pulse Rest_Pulse Maximum_Pulse
9	5	5.9856	0.8356	0.8027	Performance Age Weight Run_Pulse Maximum_Pulse
10	6	6.0492	0.8483	0.8104	Performance RunTime Age Weight Run_Pulse Maximum_Pulse
11	6	6.1758	0.8475	0.8094	RunTime Age Weight Run_Pulse Rest_Pulse Maximum_Pulse
12	6	6.6171	0.8446	0.8057	Performance RunTime Weight Run_Pulse Rest_Pulse Maximum_Pulse
13	6	6.7111	0.8440	0.8049	Performance RunTime Age Run_Pulse Rest_Pulse Maximum_Pulse
…	…	…	…	…	…

MODEL SELECTION | STEPWISE SELECTION METHODS

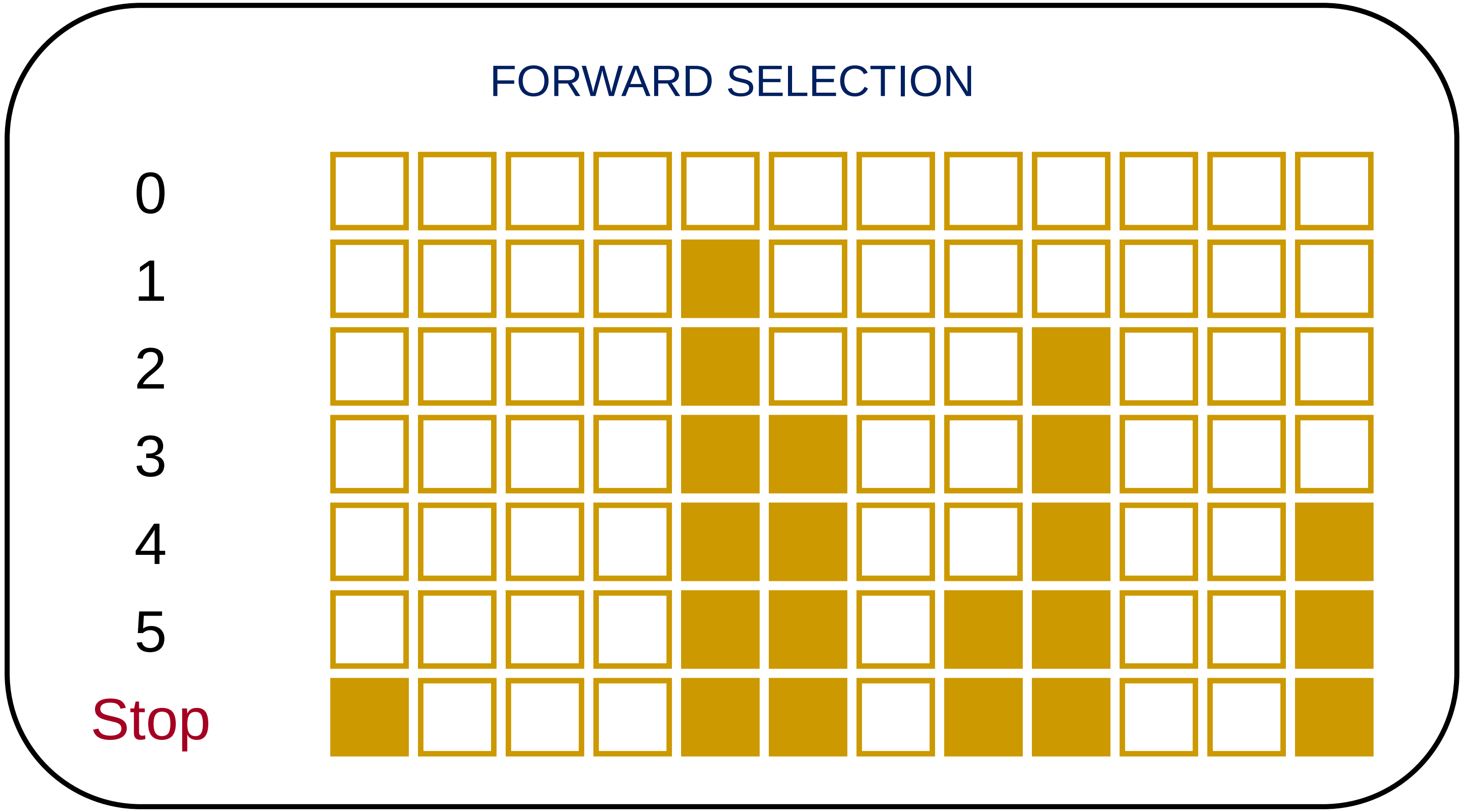


FORWARD SELECTION



BACKWARD ELIMINATION

STEPWISE SELECTION



SLENTRY=*value*
SLE=*value*

SLSTAY=*value*
SLS=*value*

STEPWISE SELECTION METHODS

```
proc reg data=sasuser.fitness plots(only)=adjrsq;
   FORWARD:  model oxygen_consumption
                   = Performance RunTime Age Weight
                     Run_Pulse Rest_Pulse Maximum_Pulse
            / selection=forward;
   BACKWARD: model oxygen_consumption
                   = Performance RunTime Age Weight
                     Run_Pulse Rest_Pulse Maximum_Pulse
            / selection=backward;
   STEPWISE: model oxygen_consumption
                   = Performance RunTime Age Weight
                     Run_Pulse Rest_Pulse Maximum_Pulse
            / selection=stepwise;
   title 'Best Models Using Stepwise Selection';
run;
quit;
```

(3) MODEL ASSUMPTION VALIDATION

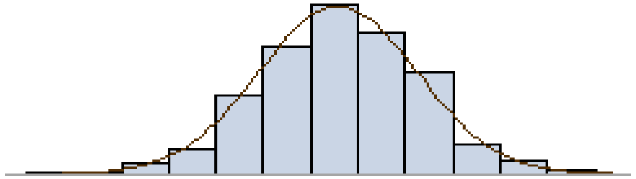
MULTIPLE LINEAR REGRESSION

MULTIPLE LINEAR REGRESSION

ASSUMPTIONS

- The mean of the Ys is accurately modeled by a linear function of the Xs.
- The assumptions for linear regression are that the error terms are independent and normally distributed with equal variance.

$\varepsilon \sim iid\ N(0,\sigma^2)$



- Therefore, evaluating model assumptions for linear regression includes checking for
  - ✓ Independent observations – независимые наблюдения
  - ✓ Normally distributed error terms – нормальность ошибки
  - ✓ Constant variance – постоянная дисперсия (по всем наблюдениям)

ASSUMPTIONS | INDEPENDENCE

- ЗНАТЬ ИСТОЧНИК ДАННЫХ: данные собранные по времени, повторные измерения, кластеризованные данные, данные экспериментов со сложными планами.
- Для данных в формате временных рядов использовать:
  - График остатков по времени или другой компоненте, определяющей порядок наблюдений
  - Статистика Durbin-Watson или автокорреляция первого порядка

WHEN THE INDEPENDENCE ASSUMPTION IS VIOLATED

Use the appropriate modeling tools to account for correlated observations:

- PROC MIXED, PROC GENMOD, or PROC GLIMMIX for repeated measures data
- PROC AUTOREG or PROC ARIMA in SAS/ETS for time-series data – *[NEXT SAS COURSE]*
- PROC SURVEYREG for survey data

ASSUMPTIONS | NORMALITY

Check that the error terms are normally distributed by examining:

- a histogram of the residuals
- a normal probability plot of the residuals
- tests for normality

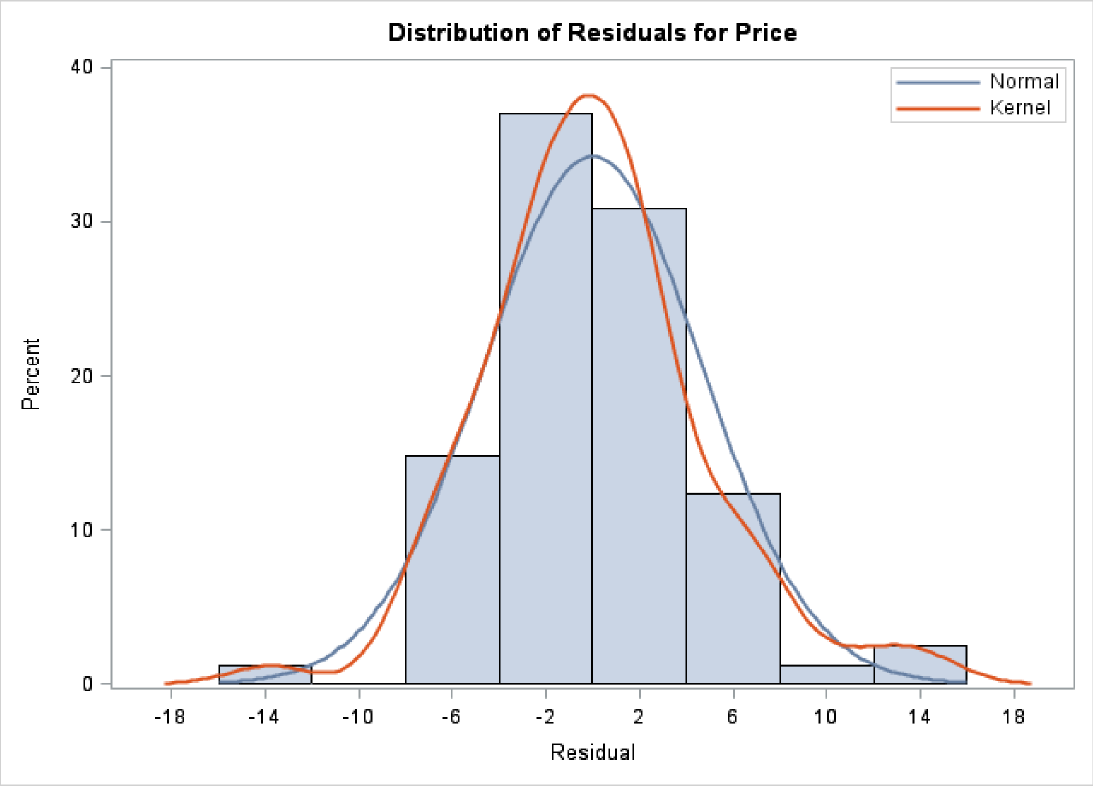
WHEN THE NORMALITY ASSUMPTION IS VIOLATED

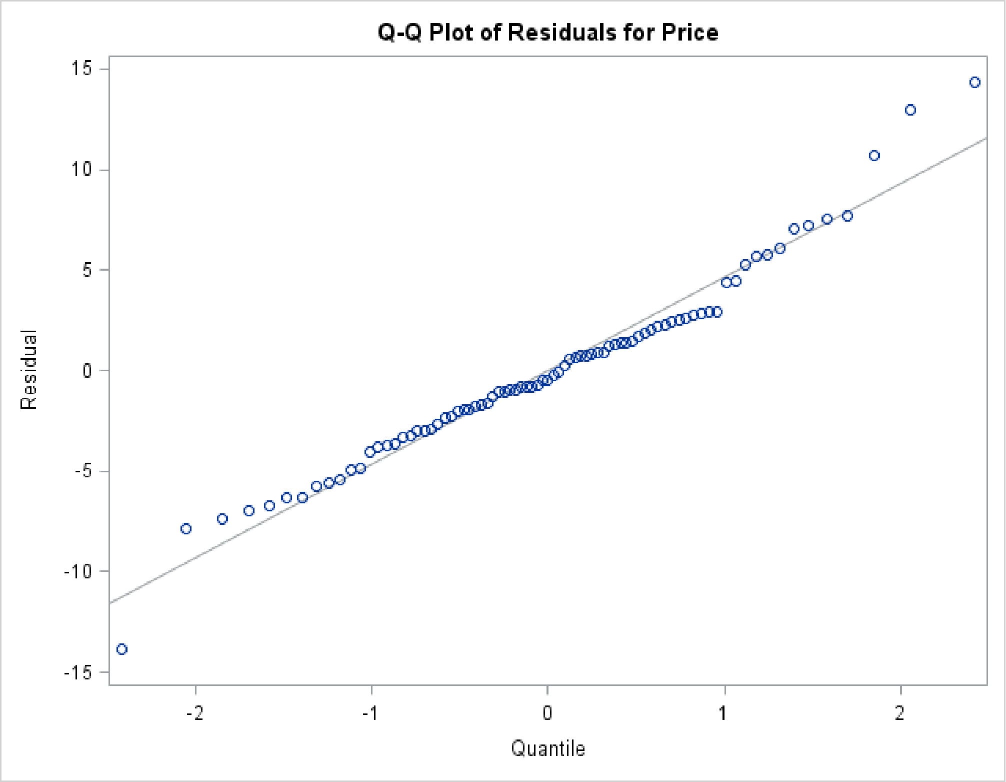
- Transform the dependent variable
- Fit a *generalized linear model* using PROC GENMOD or PROC GLIMMIX with the appropriate DIST= and LINK= option.

ASSUMPTIONS NORMALITY

```
proc reg data=sasuser.cars2  plots=all;
   model price = hwympg hwympg2 horsepower;
run;
```

Also, formal test for normality in **`proc univariate`**





ASSUMPTIONS | CONSTANT VARIANCE

Check for constant variance of the error terms by examining:

- plot of residuals versus predicted values
- plots of residuals versus the independent variables
- test for heteroscedasticity
- Spearman rank correlation coefficient between absolute values of the residuals and predicted values.

WHEN THE CONSTANT VARIANCE ASSUMPTION IS VIOLATED

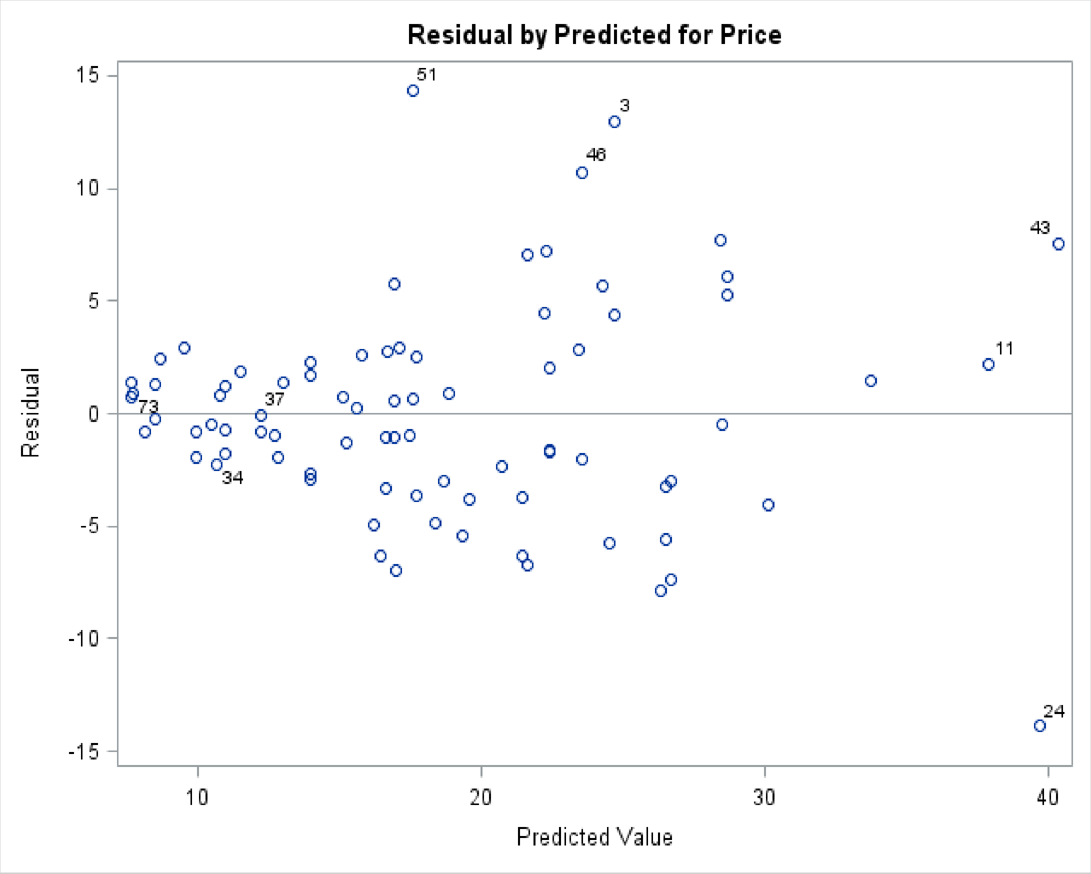
Request tests using the heteroscedasticity-consistent variance estimates.

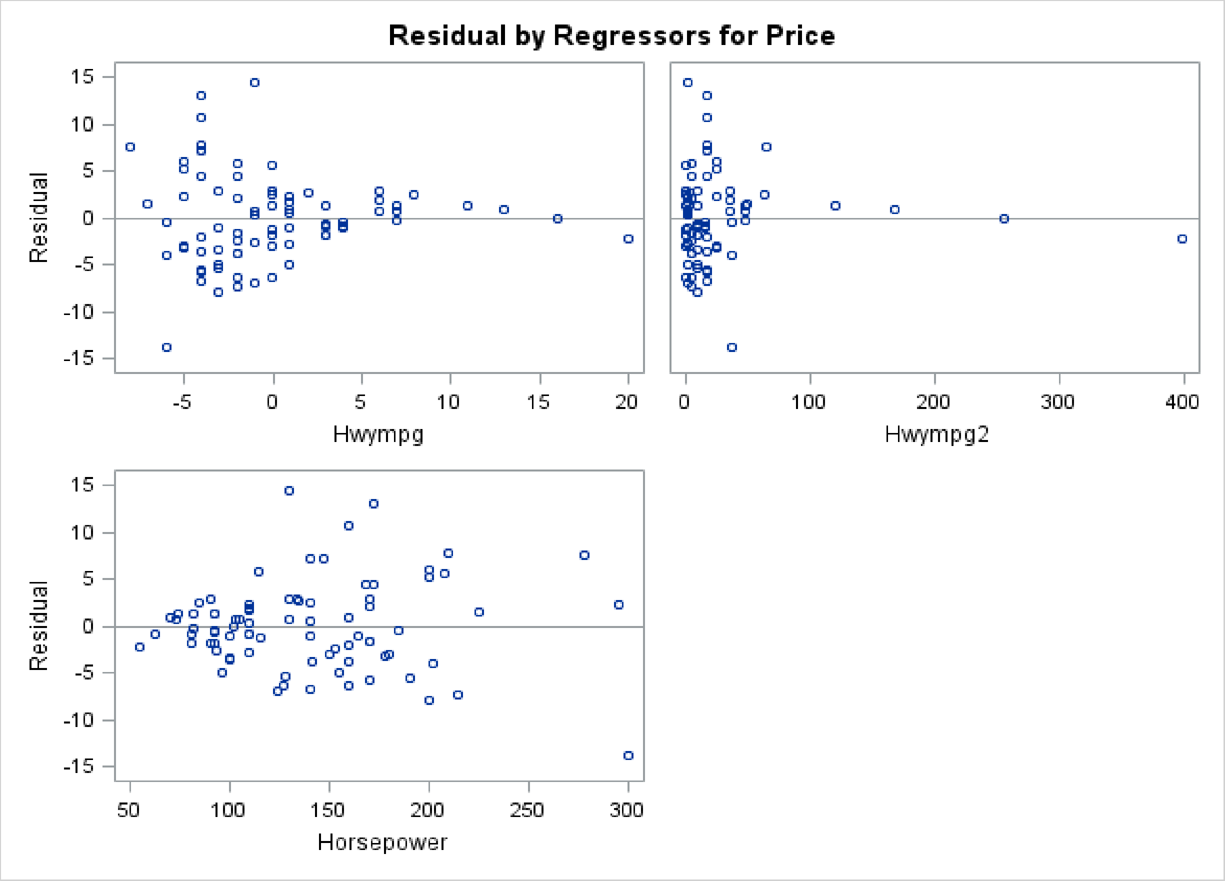
Transform the dependent variable.

Model the nonconstant variance by using:

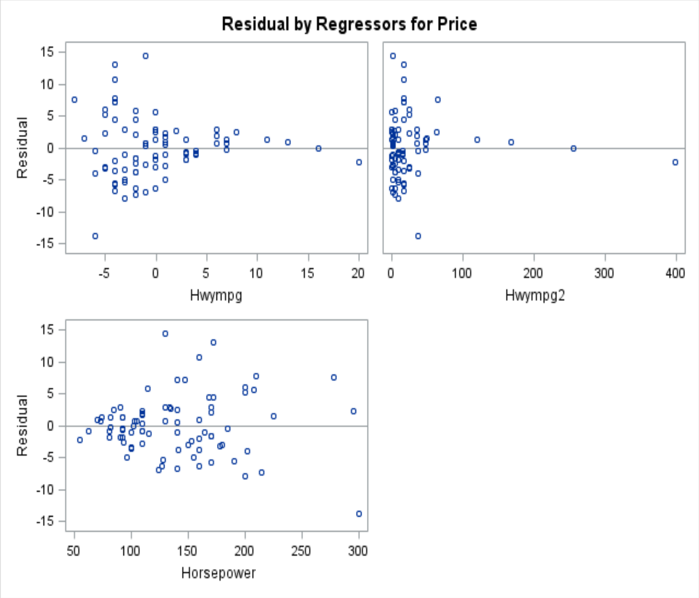
- PROC GENMOD or PROC GLIMMIX with the appropriate DIST= option
- PROC MIXED with the GROUP= option and TYPE =option
- SAS SURVEY procedures for survey data
- SAS/ETS procedures for time-series data
- Weighted least squares regression model

ASSUMPTIONS CONSTANT VARIANCE: PLOTS





ASSUMPTIONS CONSTANT VARIANCE: TESTS




```
model Y = X1 X2 X3 / white hcc hccmethod=0;
```

Parameter Estimates								
						Heteroscedasticity Consistent		
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|
Intercept	1,00	4,04	2,17	1,86	0,07	**2,68**	**1,51**	**0,14**
Hwympg	1,00	-0,80	0,21	-3,76	0,00	**0,19**	**-4,16**	**<.0001**
Hwympg2	1,00	0,04	0,01	3,04	0,00	**0,01**	**4,21**	**<.0001**
Horsepower	1,00	0,10	0,02	6,03	<.0001	**0,02**	**4,72**	**<.0001**

```
model Y = X1 X2 X3 / spec ;
```

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
8	16.49	0.0359

proc corr

[next slide …]

```
WARNING: The average covariance matrix for the SPEC test has been deemed
singular which violates an assumption of the test. Use caution when interpreting
the results of the test.
```

ASSUMPTIONS | CONSTANT VARIANCE

SPEARMAN RANK CORRELATION COEFFICIENT

- The Spearman rank correlation coefficient is available as an option in PROC CORR
- If the Spearman rank correlation coefficient between the absolute value of the residuals and the predicted values is
  - **close to zero**, then the variances are approximately equal
  - **positive**, then the variance increases as the mean increases
  - **negative**, then the variance decreases as the mean increases.

```
proc reg data=sasuser.cars2 plots (label)= all;
   model price = hwympg hwympg2 horsepower /
spec ;
   output out=check r=residual p=pred;
run;

data check;
   set check;
   abserror=abs(residual);
run;

proc corr data=check spearman nosimple;
   var abserror pred;
   title 'Spearman corr.';
run;
```

Spearman Correlation Coefficients, N = 81
Prob > |r| under H0: Rho=0

	abserror	pred
abserror	1.00000	0.60274
		<.0001
pred	0.60274	1.00000
Predicted Value of Price	<.0001	

ASSUMPTIONS LINEAR RELATION BETWEEN E[Y] AND X

Use the diagnostic plots available via the ODS Graphics output of PROC REG to evaluate the model fit:

- Plots of residuals and studentized residuals versus predicted values
- "Residual-Fit Spread" (or R-F) plot
- Plots of the observed values versus the predicted values
- Partial regression leverage plots

and…

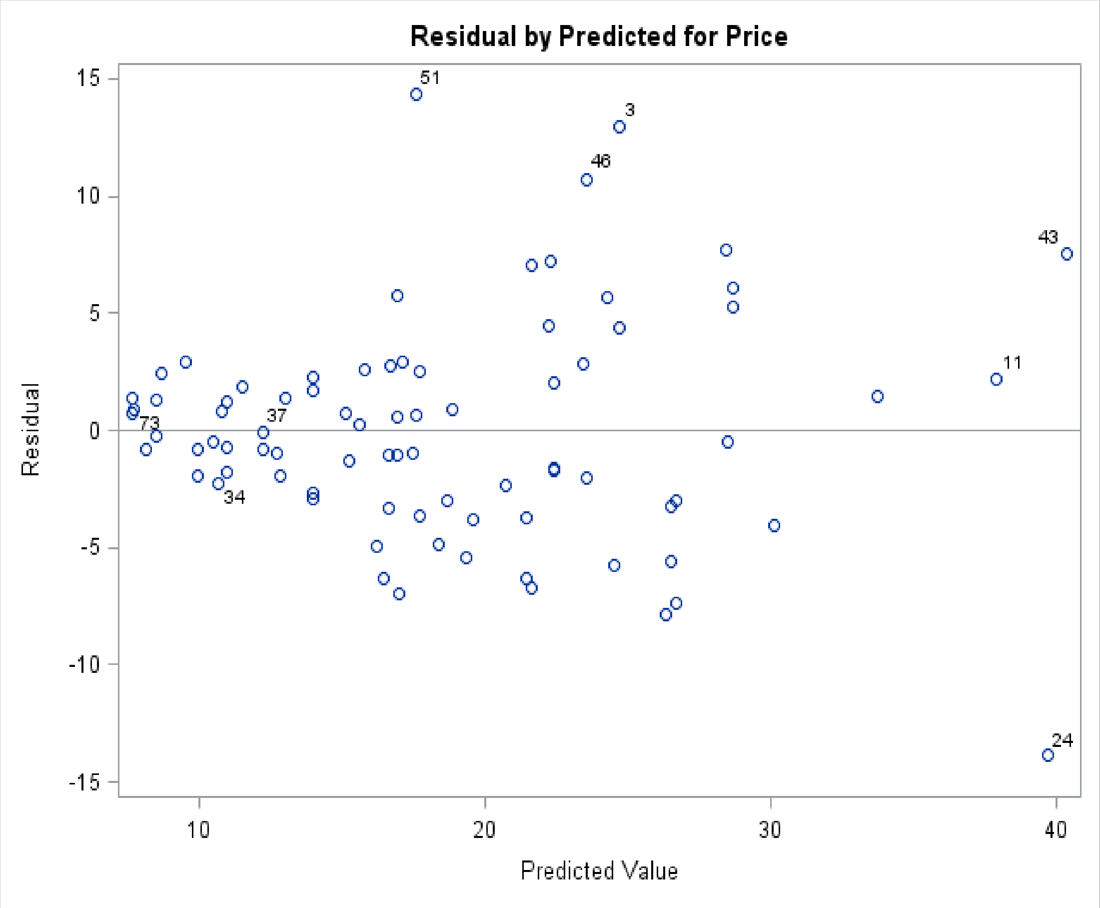
WHEN A STRAIGHT LINE IS INAPPROPRIATE

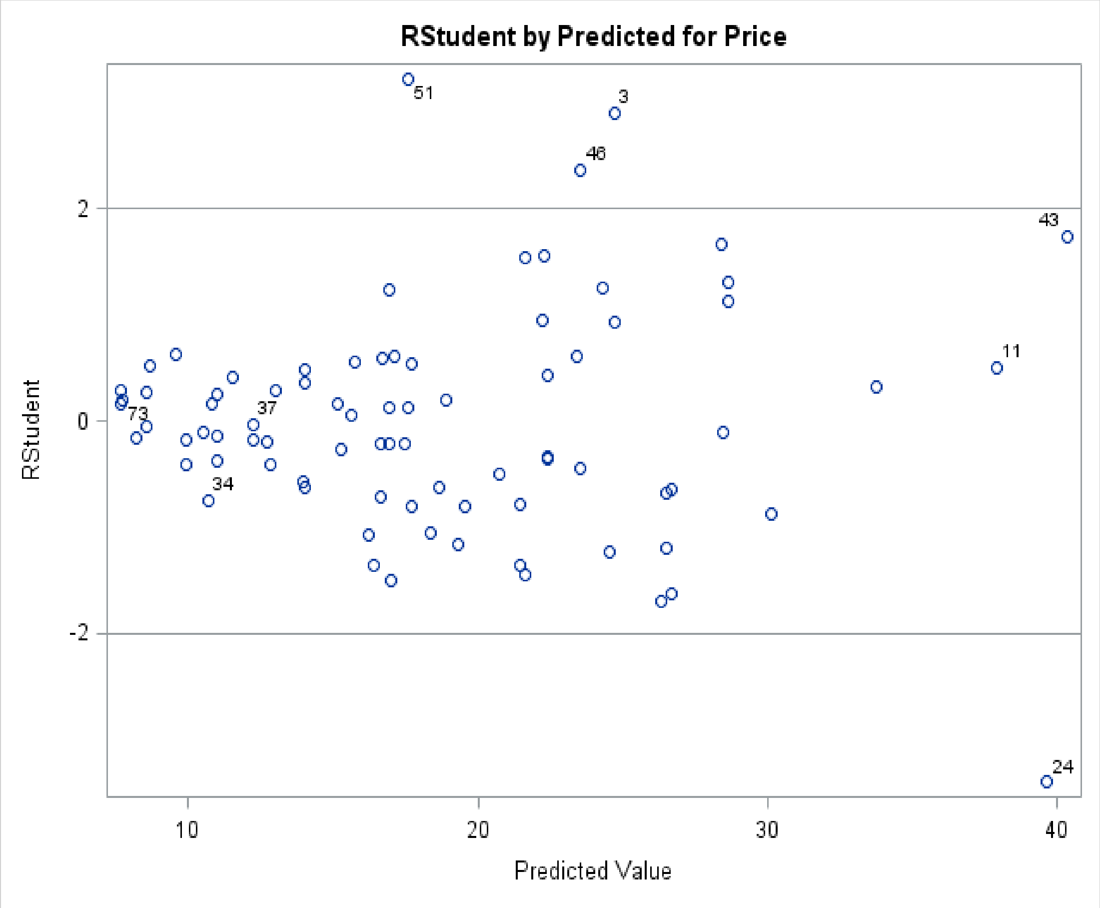
- Fit a polynomial regression model.
- Transform the independent variables to obtain linearity.
- Fit a nonlinear regression model using PROC NLIN if appropriate.
- Fit a nonparametric regression model using PROC LOESS.

- Examine model-fitting statistics such as R2, adjusted R2, AIC, SBC, and Mallows' Cp.
- Use the LACKFIT option in the MODEL statement in PROC REG to test for lack-of-fit for models that have replicates for each value of the combination of the independent variables.

ASSUMPTIONS LINEAR RELATION BETWEEN E[Y] AND X: PLOTS

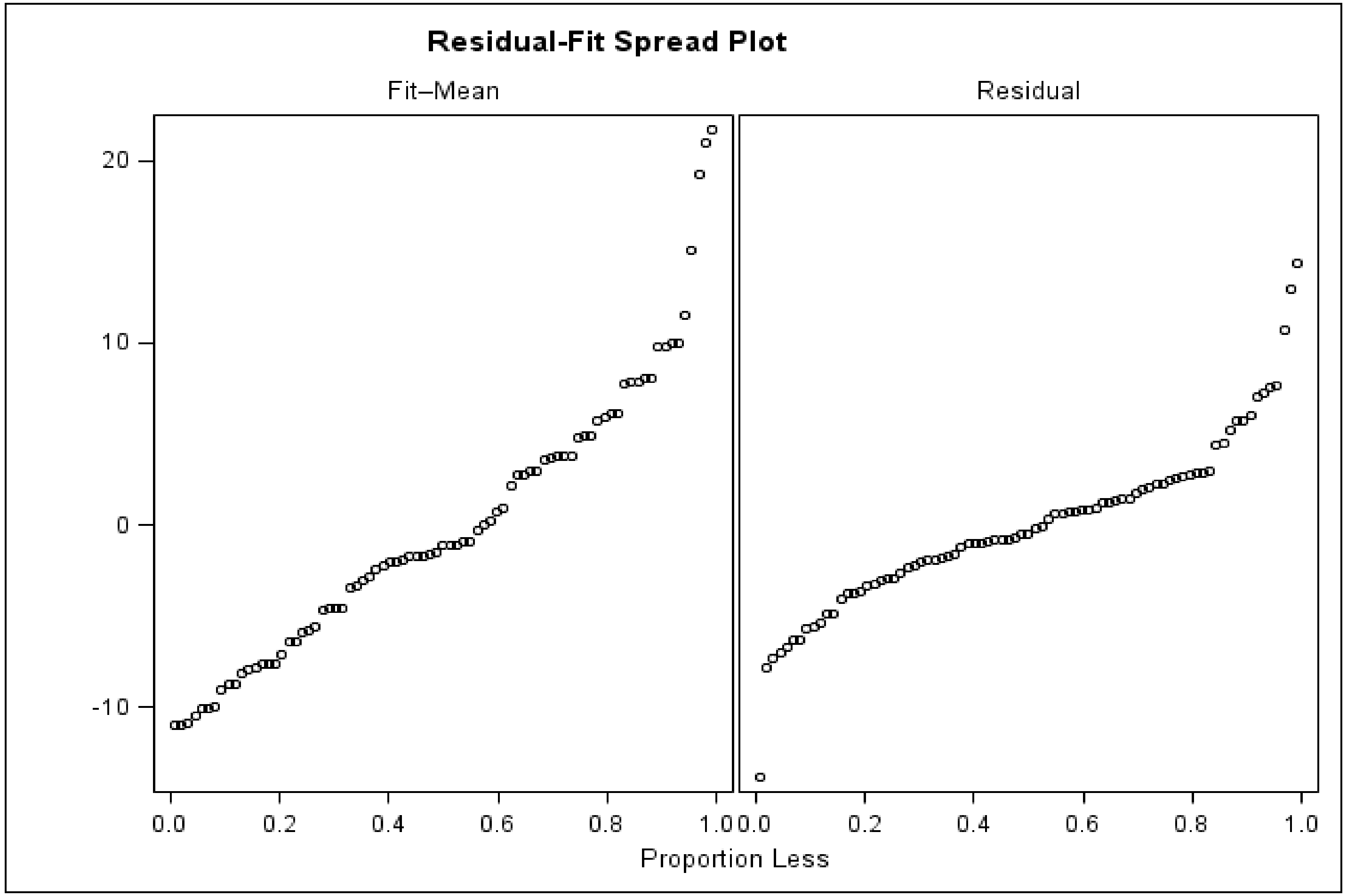
Plots of residuals and studentized residuals versus predicted values





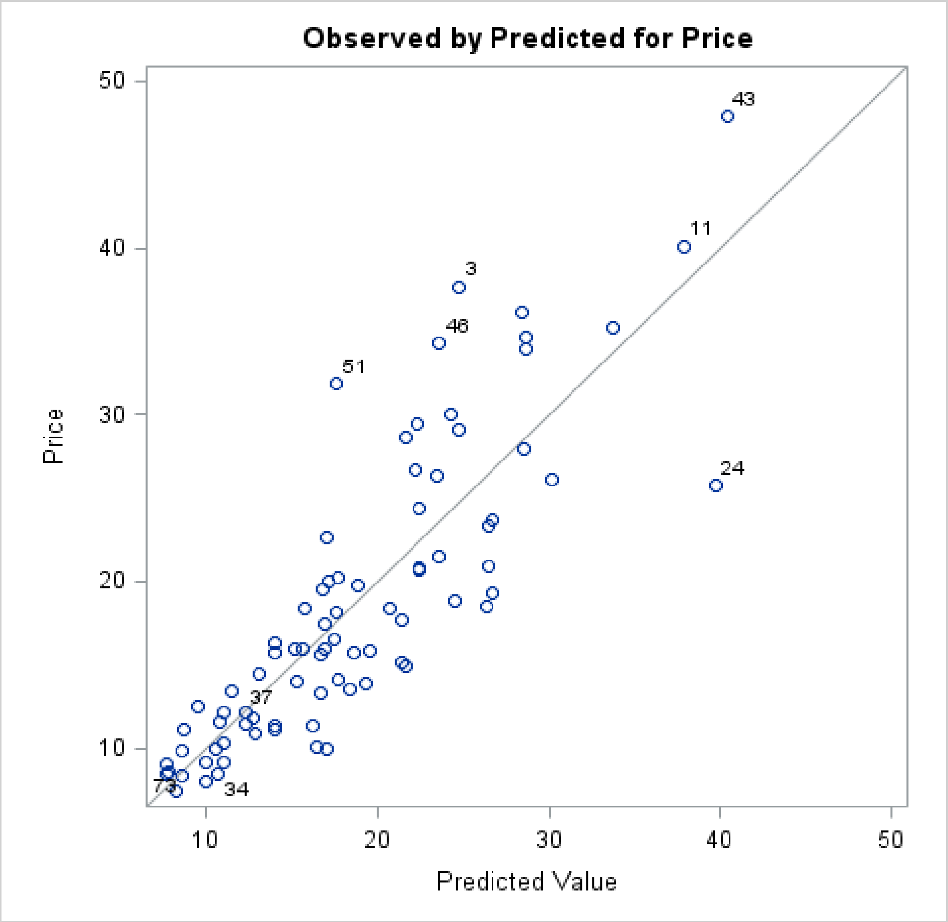
ASSUMPTIONS | LINEAR RELATION BETWEEN E[Y] AND X: PLOTS

"Residual-Fit Spread" (or R-F) plot



ASSUMPTIONS LINEAR RELATION BETWEEN E[Y] AND X: PLOTS

Plots of the observed values versus the predicted values

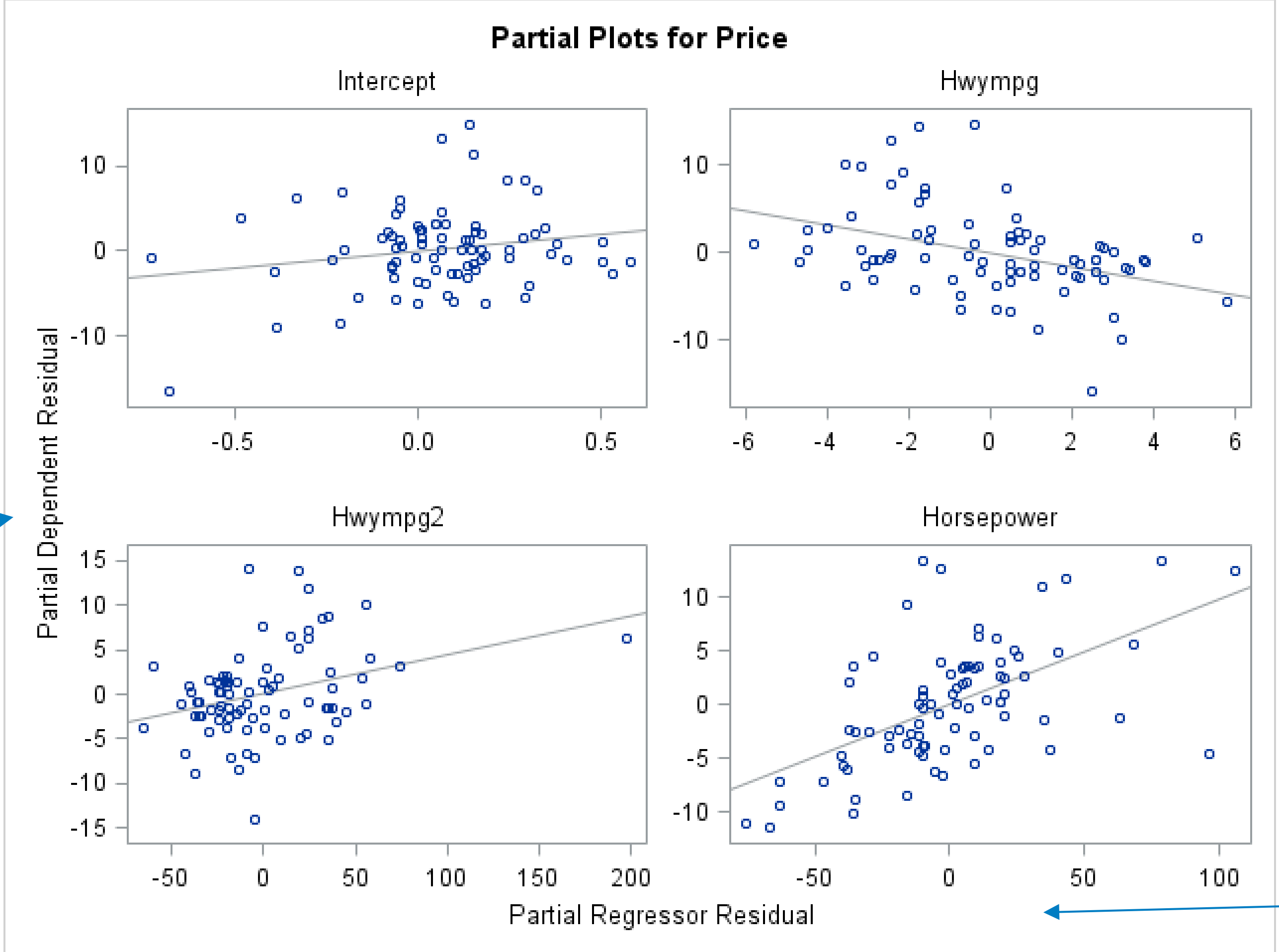


ASSUMPTIONS | LINEAR RELATION BETWEEN E[Y] AND X

Partial regression leverage plots

```
model … / partial
```

residuals for the dependent variable are calculated with the selected regressor omitted



residuals for the selected regressor are calculated from a model where the selected regressor is regressed on the remaining regressors

(4) COLLINEARITY AND INFLUENTIAL OBSERVATION DETECTION

MULTIPLE LINEAR REGRESSION

WHAT ELSE CAN HAPPEN…

MULTICOLLINEARITY

ПОЛЕЗНЫЙ ПРИМЕР НА SUPPORT.SAS.COM:
http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_reg_sect038.htm

СПОСОБЫ ОБНАРУЖЕНИЯ:

- Correlation statistics (PROC CORR)
- Variance inflation factors (VIF option in the MODEL statement in PROC REG)
- Condition index values (COLLIN and COLLINOINT options in the MODEL statement in PROC REG)

WHEN THERE IS MULTICOLLINEARITY

- Exclude redundant independent variables.
- Use biased regression techniques such as ridge regression or principal component regression.
- Center the independent variables in polynomial regression models.
- PROC VARCLASS to select vars *[next time]*

ПРОБЛЕМЫ:

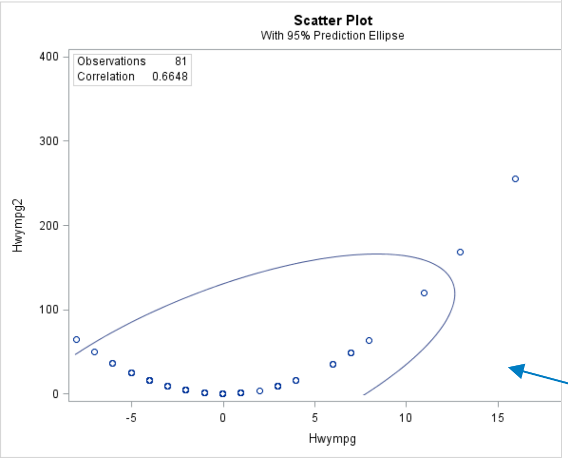
- Некорректный результат пошаговых методов выбора переменных
- Некорректная оценка значений коэффициентов модели: очень большие/маленькие значения, неверный знак

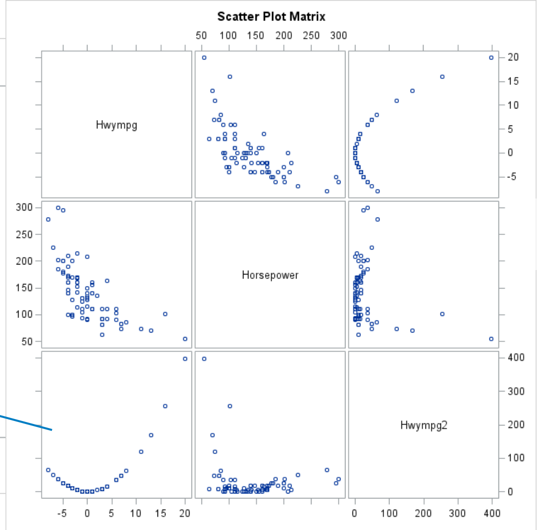
- опасны, когда цель моделирования – исследование
- не очень важны, когда цель модели – предсказание (*однако может снизиться устойчивость модели*)

WHAT ELSE CAN HAPPEN…

MULTICOLLINEARITY

```
proc reg data=sasuser.cars2
         plots (label)=all;
model price = hwympg
              hwympg2
              horsepower
    / vif collin collinoint;
run;
```





Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1,00	4,04	2,17	1,86	0,07	0,00
Hwympg	1,00	-0,80	0,21	-3,76	0,00	4,07
Hwympg2	1,00	0,04	0,01	3,04	0,00	2,27
Horsepower	1,00	0,10	0,02	6,03	<.0001	2,37

> 4- 10 – плохо

Collinearity Diagnostics						
Number	Eigenvalue	Condition Index	Proportion of Variation			
			Intercept	Hwympg	Hwympg2	Horsepower
1,00	2,18	1,00	0,01	0,00	0,03	0,01
2,00	1,53	1,19	0,00	0,09	0,07	0,00
3,00	0,27	2,85	0,03	0,32	0,69	0,00
4,00	0,03	9,25	0,96	0,58	0,21	0,99

не интересен

> 0,5

> 0,5

В этой таблице свободный член не используется в расчетах

Около 10 – плохо, 100 – совсем лохо

Collinearity Diagnostics (intercept adjusted)					
Number	Eigenvalue	Condition Index	Proportion of Variation		
			Hwympg	Hwympg2	Horsepower
1,00	2,06	1,00	0,05	0,06	0,06
2,00	0,80	1,61	0,00	0,28	0,26
3,00	0,14	3,79	0,95	0,66	0,68

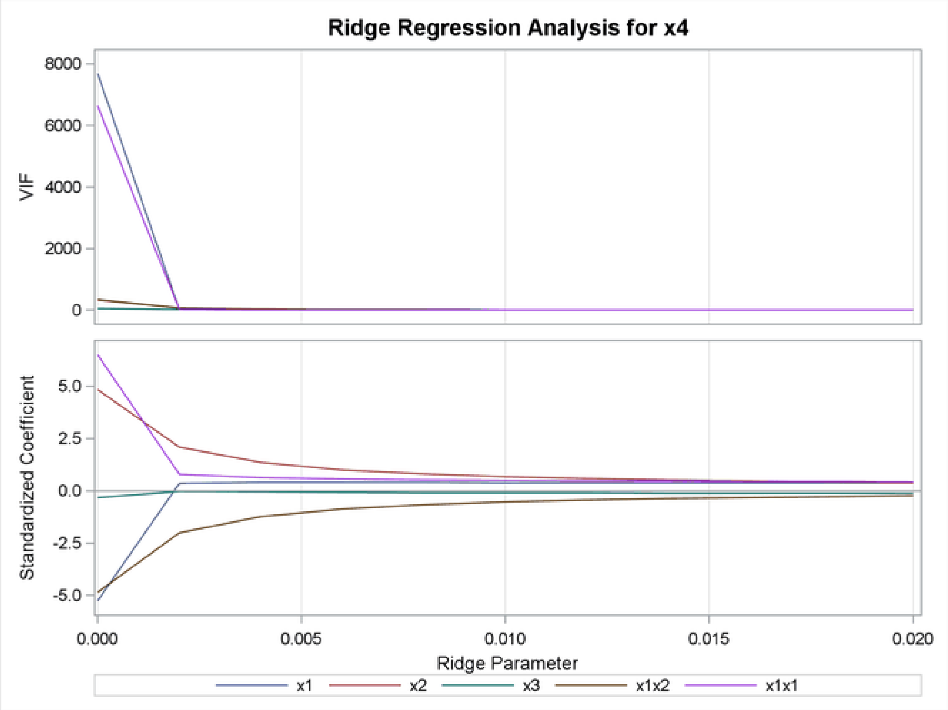
WHAT ELSE CAN HAPPEN…

MULTICOLLINEARITY: RIDGE REG

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \Big)^2, \quad \text{subject to } \sum_{j=1}^{p} \beta_j^2 \le t,$$

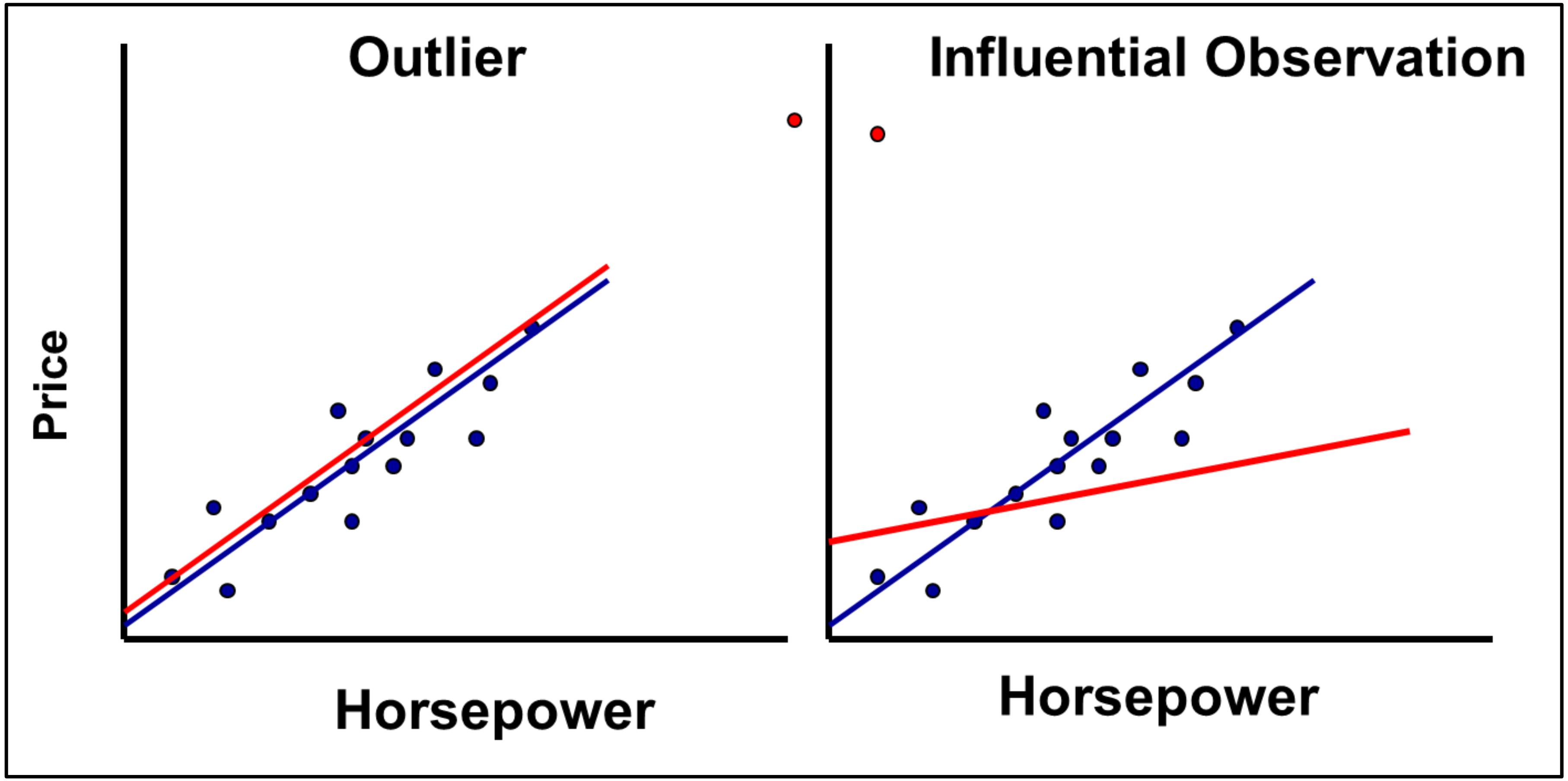
$$\hat{\beta}^{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

```
proc reg data=acetyl outvif
         outest=b ridge=0 to 0.02 by .002;
   model x4=x1 x2 x3 x1x2 x1x1;
run;
```



WHAT ELSE CAN HAPPEN…

INFLUENTIAL OBSERVATIONS



WHAT ELSE CAN HAPPEN…

INFLUENTIAL OBSERVATIONS

$$h_i = (\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)_{ii}$$

RSTUDENT residual

measures the change in the residuals when an observation is deleted from the model.

$$RSTUDENT = \frac{r_i}{s_{(i)}\sqrt{1-h_i}}$$

Leverage

measures how far an observation is from the cloud of observed data points

Cook's D

measures the simultaneous change in the parameter estimates when an observation is deleted.

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p \text{ MSE}}.$$

DFFITS

measures the change in predicted values when an observation is deleted from the model.

$$DFFITS = \frac{\overline{y}_i - \overline{y_{(i)}}}{s_{(i)}\sqrt{h(i)}}$$

(…continued …)

WHAT ELSE CAN HAPPEN…

INFLUENTIAL OBSERVATIONS

$$DFBETA_{j(i)} = \frac{b_j - b_{j(i)}}{\hat{\sigma}(b_j)}$$

DFBETAs

measures the change in each parameter estimate when an observation is deleted from the model.

$$COVRATIO_i = \frac{\left| s_{(i)}^2 \left(X_{(i)}' X_{(i)} \right)^{-1} \right|}{\left| s^2 \left(X'X \right)^{-1} \right|}$$

COVRATIO

measures the change in the precision of the parameter estimates when an observation is deleted from the model

WHEN THERE ARE INFLUENTIAL OBSERVATIONS

- Make sure that there are no data errors.
- Perform a sensitivity analysis and report results from different scenarios.
- Investigate the cause of the influential observations and redefine the model if appropriate.
- Delete the influential observations if appropriate and document the situation.
- Limit the influence of outliers by performing robust regression analysis using PROC ROBUSTREG.

INFLUENTIAL OBSERVATIONS

IDENTIFYING INFLUENTIAL OBSERVATIONS – SUMMARY OF SUGGESTED CUTOFFS

Influential Statistics	Cutoff Values
RSTUDENT Residuals	|RSTUDENT| > 2
LEVERAGE	LEVERAGE > $\frac{2p}{n}$
Cook's D	CooksD > $\frac{4}{n}$
DFFITS	|DFFITS| > $2\sqrt{\frac{p}{n}}$
DFBETAS	|DFBETAS| > $\frac{2}{\sqrt{n}}$
COVRATIO	COVRATIO < $1-\frac{3p}{n}$ or COVRATIO > $1+\frac{3p}{n}$

support.sas.com on proc reg

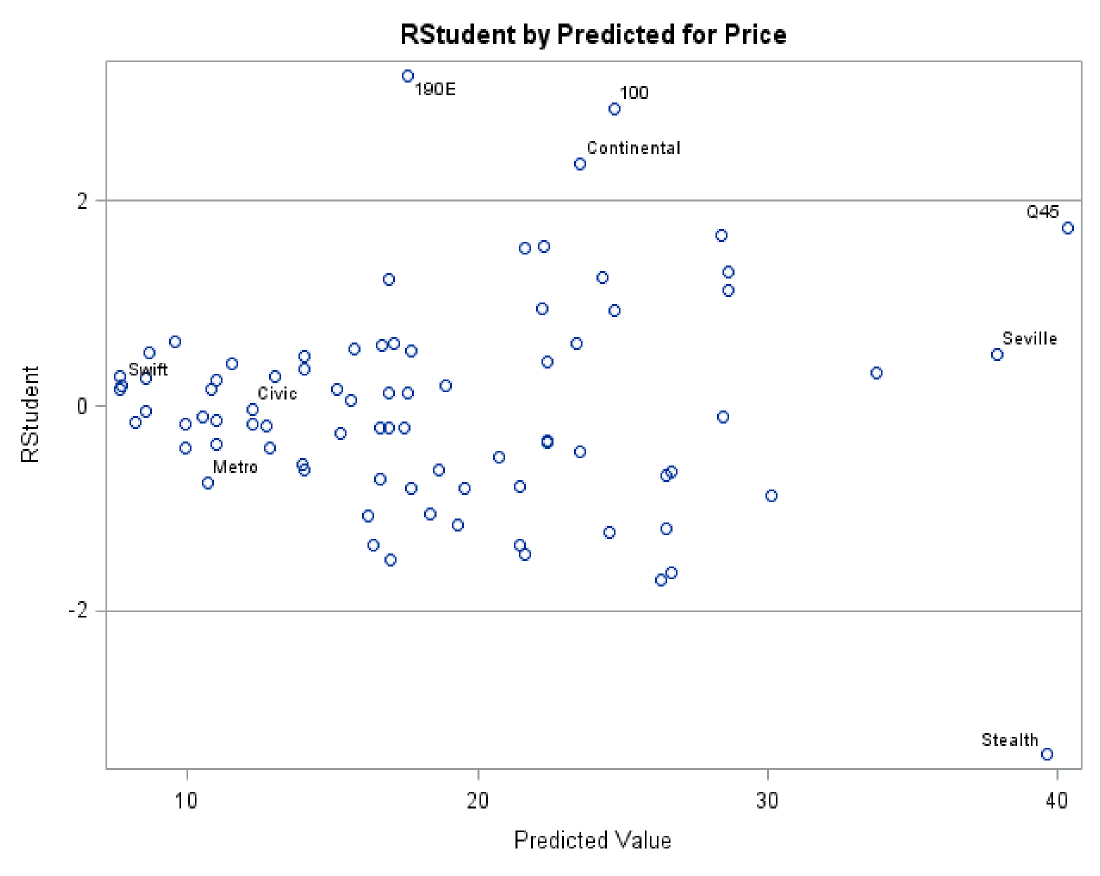
INFLUENTIAL OBSERVATIONS

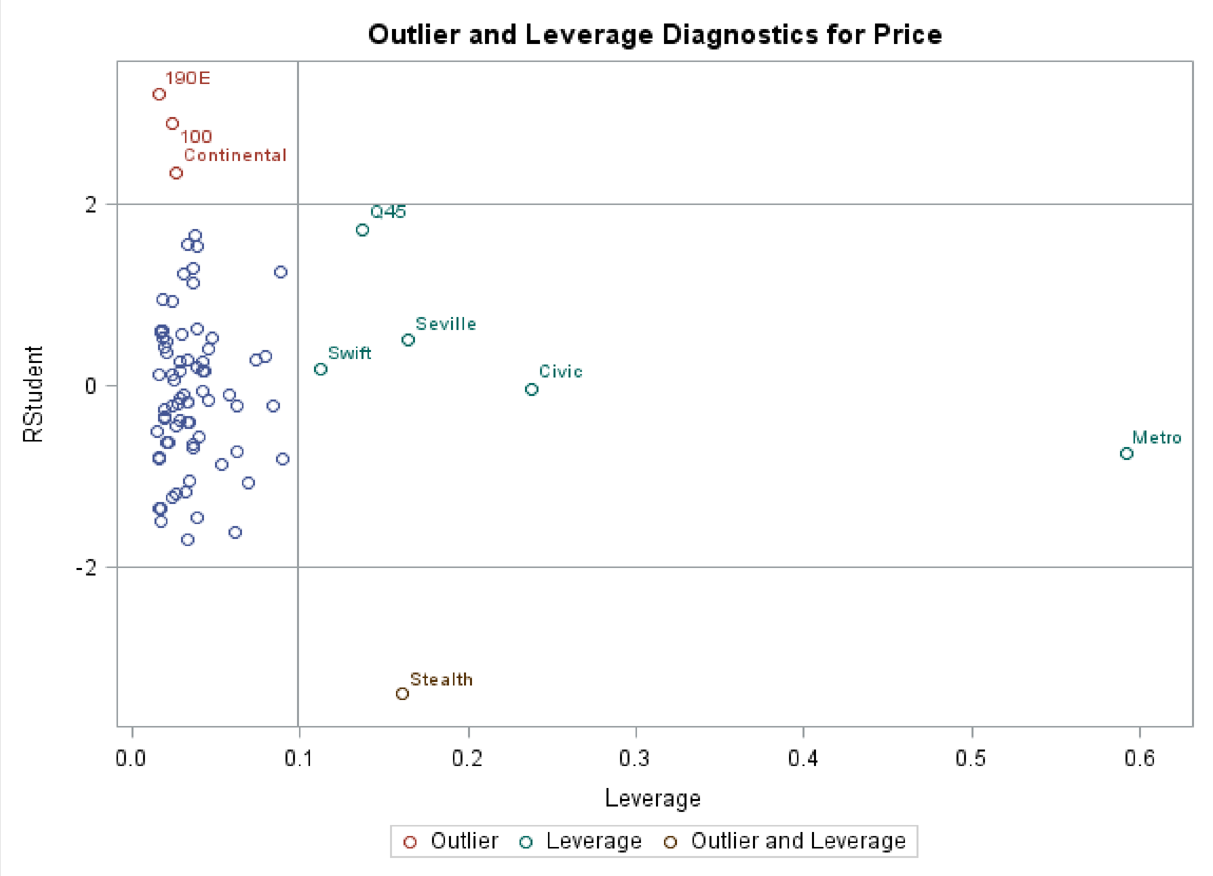
CODE

```
proc reg data=sasuser.cars2  plots (label)=all;
   model price = hwympg hwympg2 horsepower
    / influence;
   id model;
   output out=check r=residual p=pred h=leverage rstudent=rstudent covratio=CVR;
   plot COVRATIO.* (hwympg hwympg2 horsepower) / vref=(0.88 1.11) ;
run;

%let numparms = 4; %let numobs = 81;
data influence;
 set check;
 absrstud=abs(rstudent);
 if absrstud ge 2 then output;
 else if leverage ge (2*&numparms /&numobs) then output;
run;
proc print data=influence;
 var manufacturer model price hwympg horsepower;
run;
```

INFLUENTIAL OBSERVATIONS

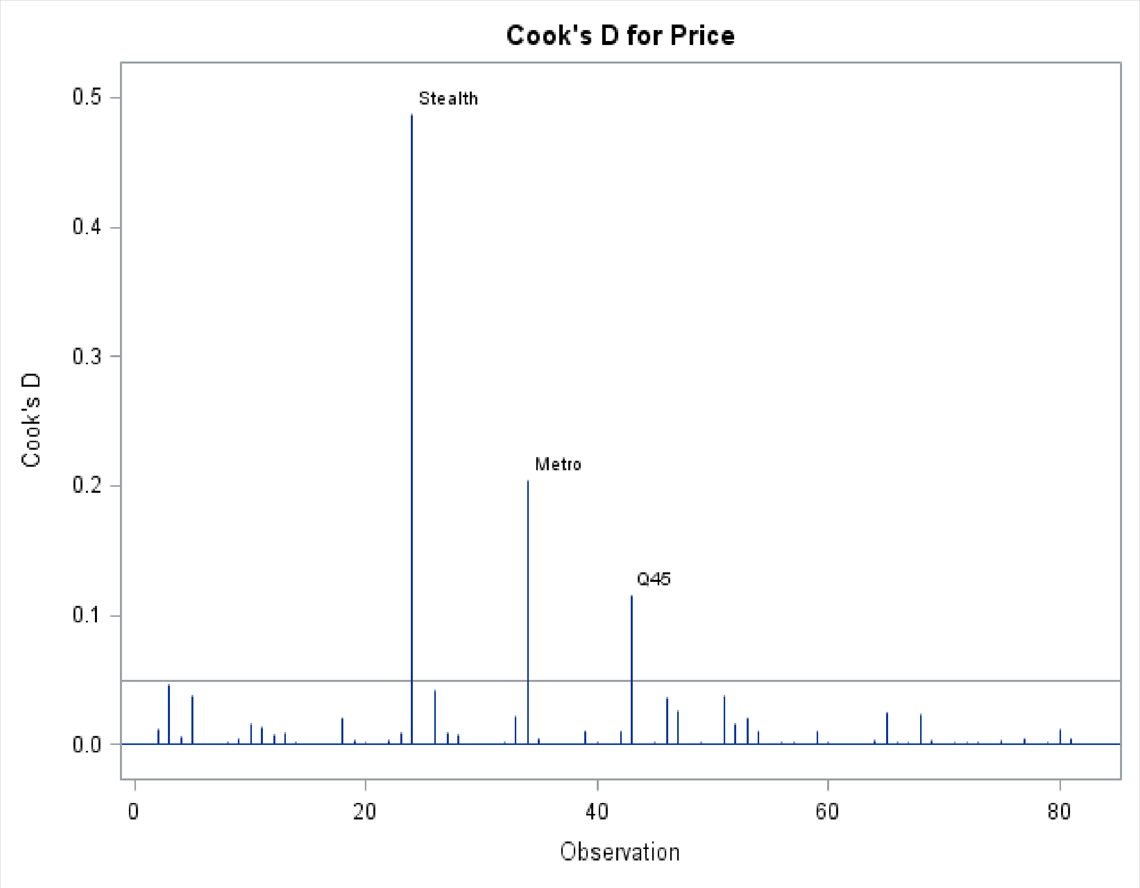
PLOTS: RSTUDENT & LEVERAGE

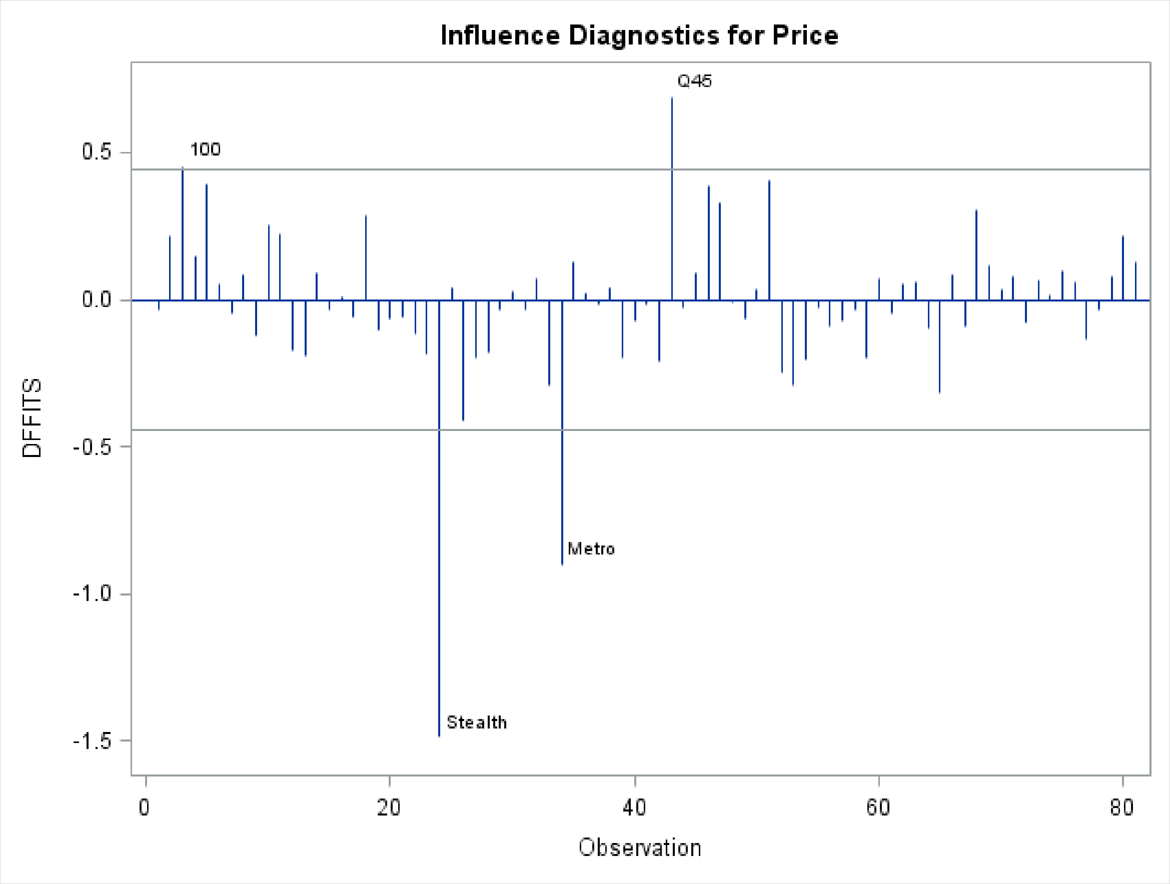




INFLUENTIAL OBSERVATIONS

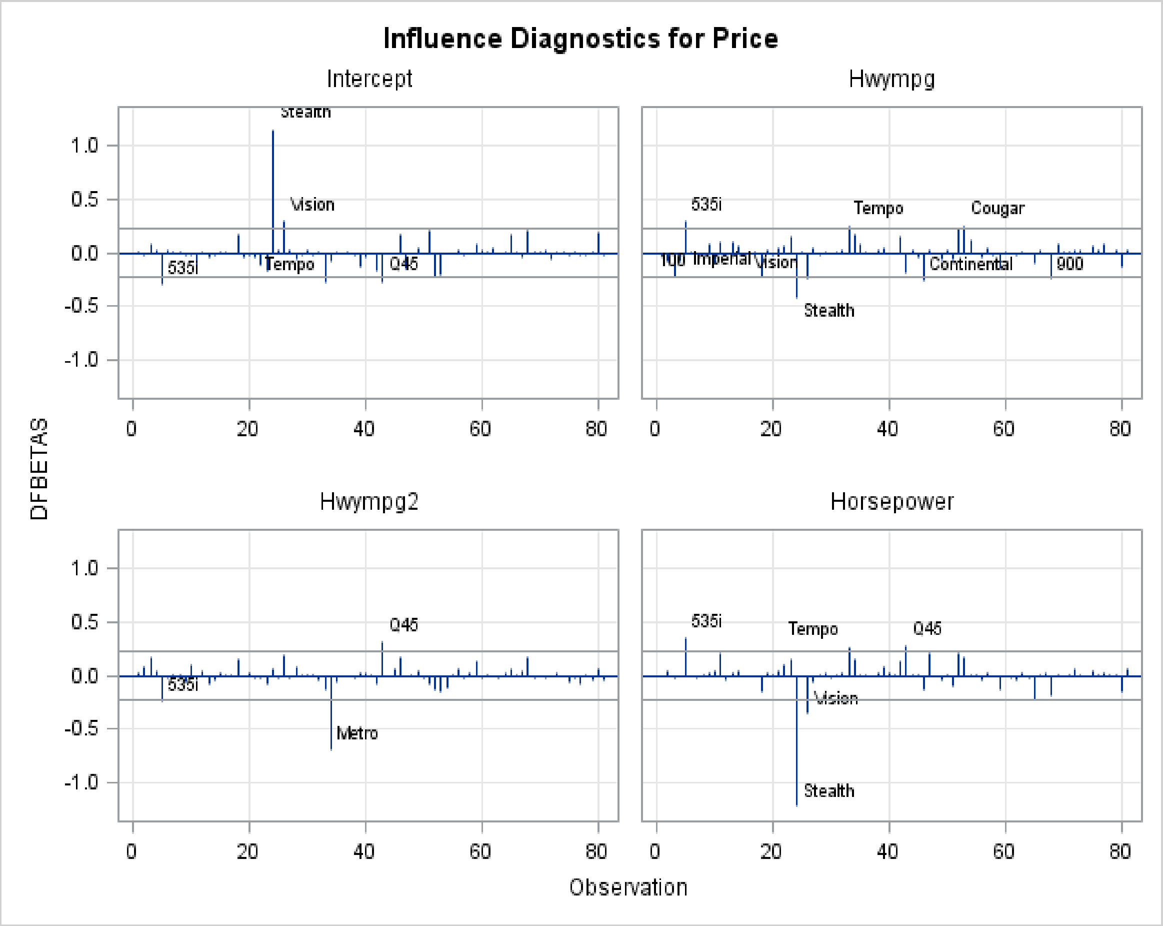
PLOTS: COOK'S D & DFFITS

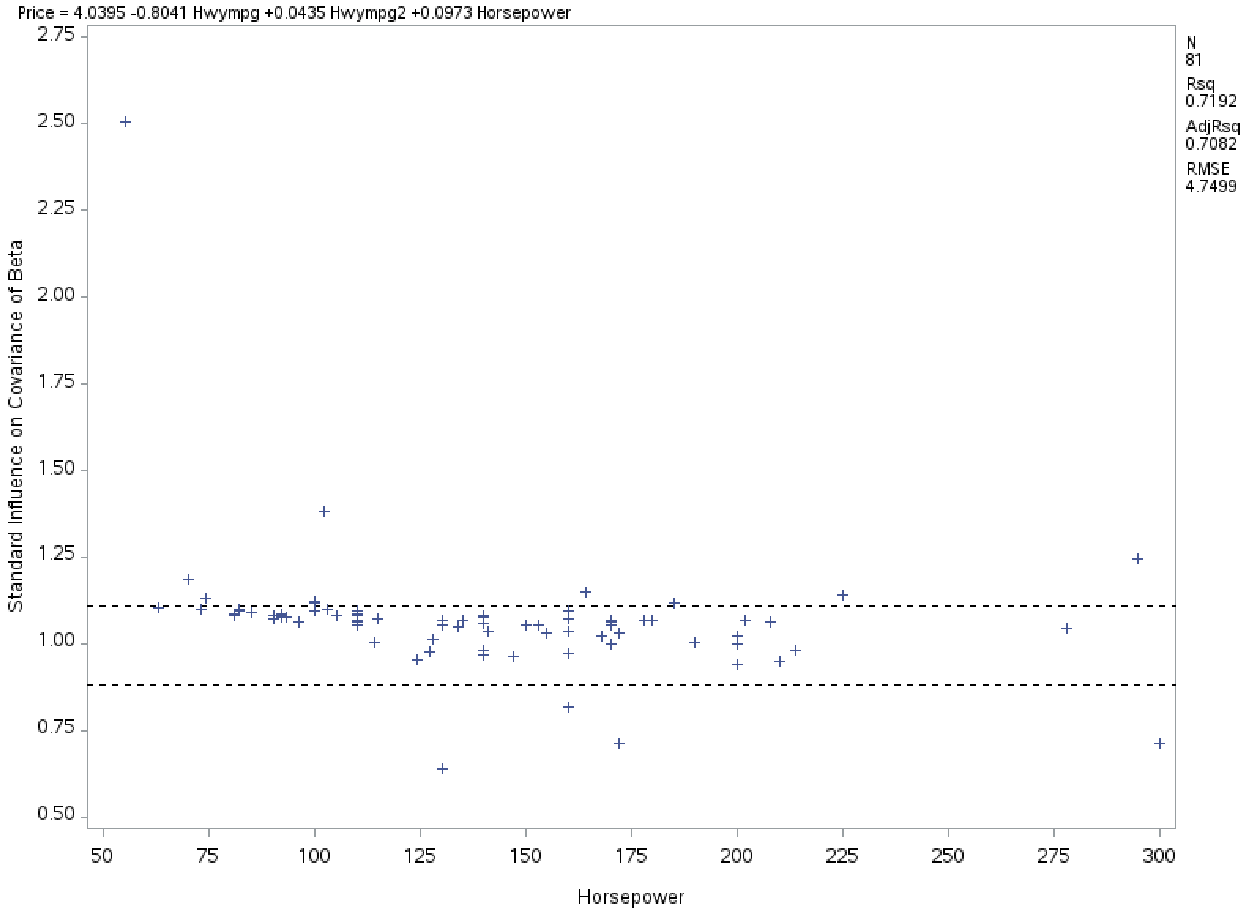




INFLUENTIAL OBSERVATIONS

PLOTS: DFBETAS & COVRATIO





INFLUENTIAL OBSERVATIONS

CODE

```
proc reg data=sasuser.cars2  plots (label)=all;
   model price = hwympg hwympg2 horsepower
    / influence;
   id model;
   output out=check r=residual p=pred h=leverage rstudent=rstudent covratio=CVR;
   plot COVRATIO.* (hwympg hwympg2 horsepower) / vref=(0.88 1.11) ;
run;

%let numparms = 4; %let numobs = 81;
data influence;
 set check;
 absrstud=abs(rstudent);
 if absrstud ge 2 then output;
 else if leverage ge (2*&numparms /&numobs) then output;
run;
proc print data=influence;
 var manufacturer model price hwympg horsepower;
run;
```

HOME WORK

- Same as at lecture
- POLYNOMIAL REGRESSION
- PROC GLMSELECT
- BOX-COX ETC. TRANSFORMATION