DIAGNOSTICS & SUBSET SELECTION

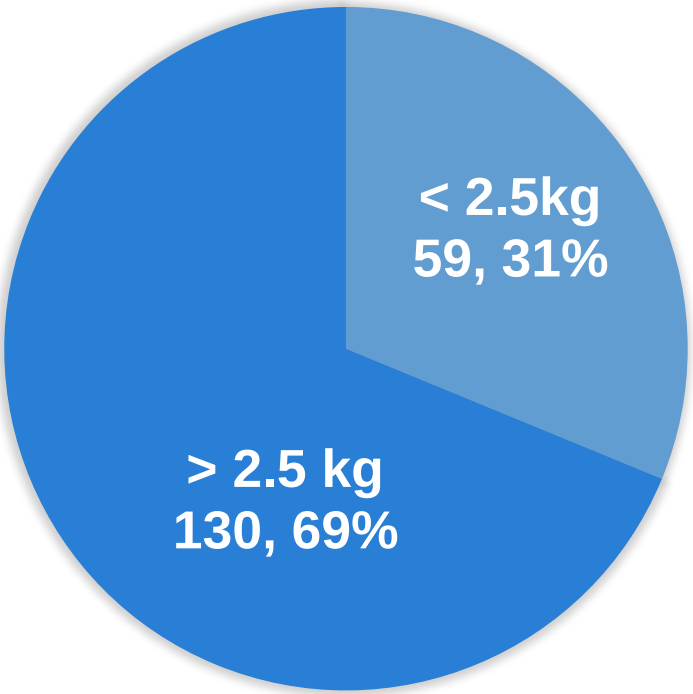
Dr. Aric LaBarr

Institute for Advanced Analytics

SUBSET SELECTION METHODS

Birth Weight Data Set

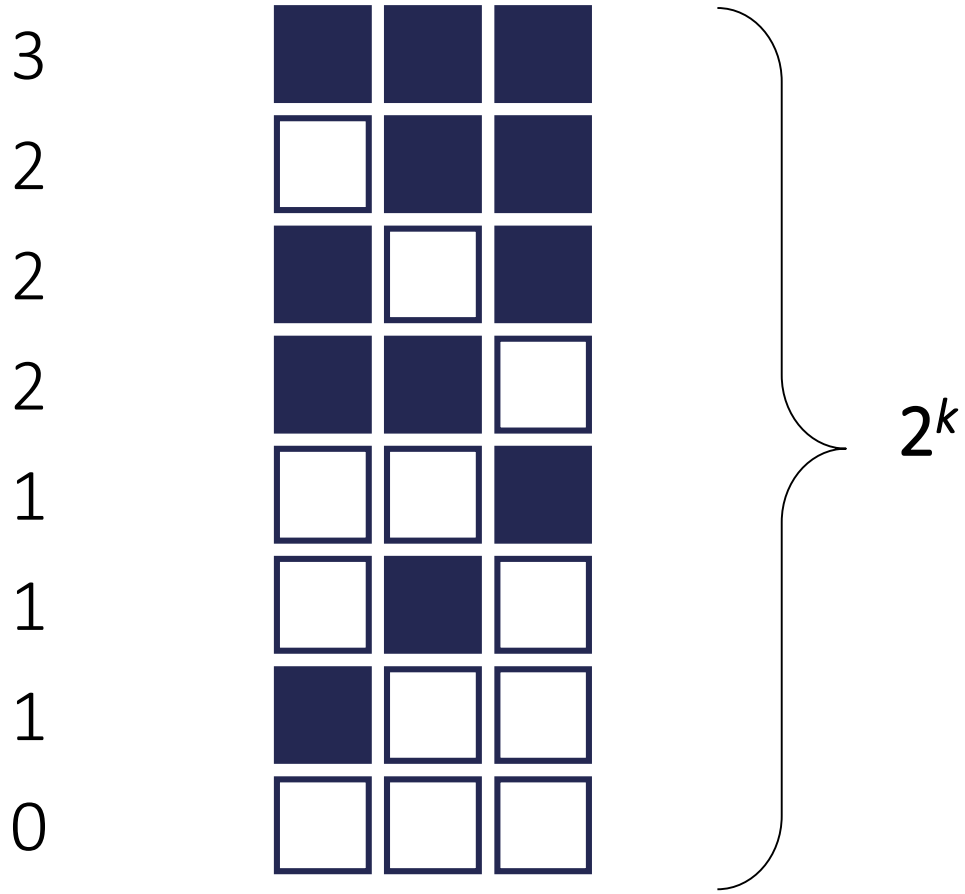
- Model the association between various factors and child being born with low birth weight (< 2.5kg)
- 189 observations in the data set



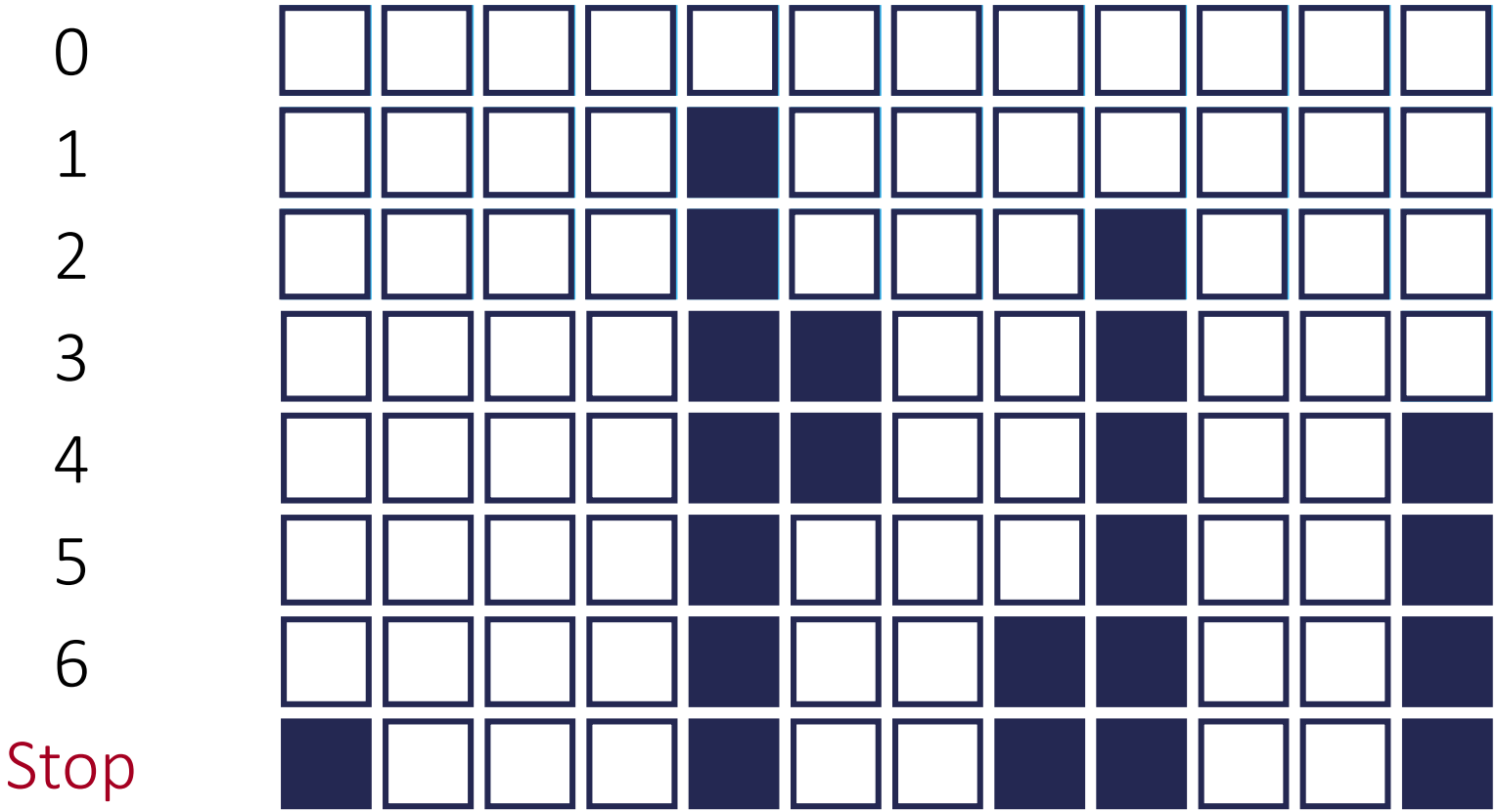
Birth Weight Data Set

- Model the association between various factors and child being born with low birth weight (< 2.5kg)
- Predictors:
  - **age:** mother's age (years)
  - **lwt:** mother's weight at last menstrual period (lbs)
  - **smoke:** mother's smoking status during pregnancy
  - **race:** mother's race (1=White, 2 = Black, 3 = Other)
  - **ptl:** number of premature labors
  - **ht:** history of hypertension
  - **ui:** uterine irritability
  - **ftv:** number of physician visits during first trimester

Best Subsets



Stepwise Selection



Stepwise Selection – SAS

```
proc logistic data=logistic.lowbwt plots(only)=(oddsratio);
   class race(ref='white') / param=ref;
   model low(event='1') = age race lwt smoke /
                          selection=stepwise
                          slentry=0.03 slstay=0.03
                          clodds=pl;
   title 'Modeling Low Birth Weight';
run;
quit;
```

Stepwise Selection – SAS

Note: No (additional) effects met the 0.03 significance level for entry into the model.

Summary of Stepwise Selection							
Step	Effect		DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > Chi Sq
	Entered	Removed					
1	lwt		1	1	5.4382		0.0197

Stepwise Selection – SAS

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
lwt	1	5.1921	0.0227

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.9983	0.7853	1.6161	0.2036
lwt	1	-0.0141	0.00617	5.1921	0.0227

Stepwise Selection – R

```
full.model <- glm(low ~ age + lwt + factor(smoke) + factor(race),
                  data = bwt, family = binomial(link = "logit"))

empty.model <- glm(low ~ 1, data = bwt,
                   family = binomial(link = "logit"))

step.model <- step(empty.model,
                   scope = list(lower=formula(empty.model),
                                upper=formula(full.model)),
                   direction = "both")
```

Stepwise Selection – R

```
## Start:  AIC=236.67
## low ~ 1
##
##                  Df Deviance    AIC
## + lwt             1   228.69 232.69
## + factor(smoke)   1   229.81 233.81
## + factor(race)    2   229.66 235.66
## + age             1   231.91 235.91
## <none>                234.67 236.67
##
## Step:  AIC=232.69
## low ~ lwt
##
##                  Df Deviance    AIC
## + factor(smoke)   1   224.34 230.34
## + factor(race)    2   223.26 231.26
## <none>                228.69 232.69
## + age             1   227.12 233.12
## - lwt             1   234.67 236.67
##
```

Stepwise Selection – R

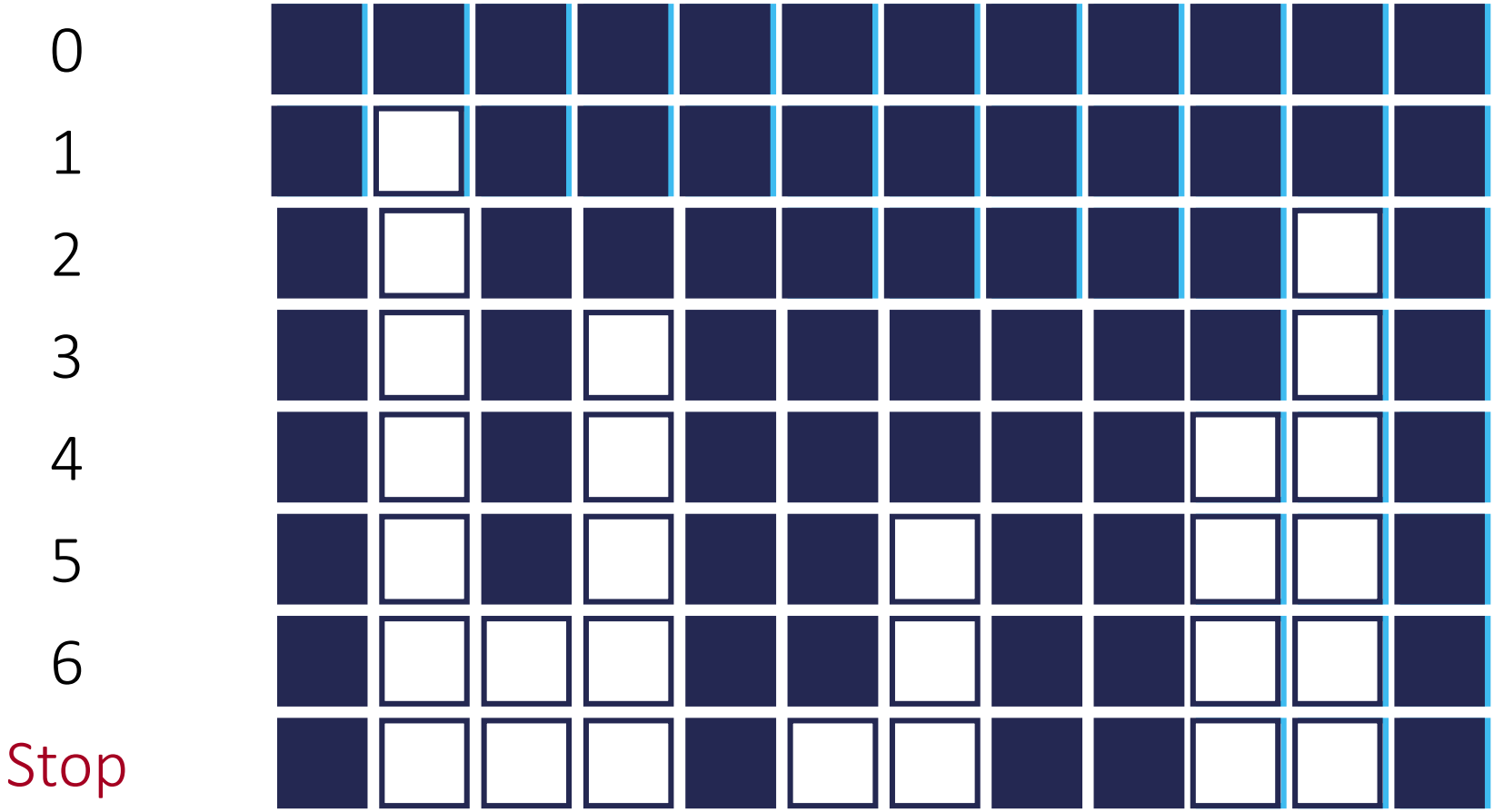
```
## Step:  AIC=230.34
## low ~ lwt + factor(smoke)
##
##                  Df Deviance    AIC
## + factor(race)    2   215.01 225.01
## <none>                224.34 230.34
## + age             1   222.88 230.88
## - factor(smoke)   1   228.69 232.69
## - lwt             1   229.81 233.81
##
## Step:  AIC=225.01
## low ~ lwt + factor(smoke) + factor(race)
##
##                  Df Deviance    AIC
## <none>                215.01 225.01
## + age             1   214.58 226.58
## - lwt             1   219.97 227.97
## - factor(race)    2   224.34 230.34
## - factor(smoke)   1   223.26 231.26
```

Stepwise Selection – R

```r
summary(step.model)
```

```
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        1.18087    1.00983   1.169  0.24225
## lwt               -0.01326    0.00631  -2.101  0.03562 *
## factor(smoke)1     1.06001    0.37832   2.802  0.00508 **
## factor(race)other -0.31958    0.52560  -0.608  0.54317
## factor(race)white -1.29009    0.51087  -2.525  0.01156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 215.01  on 184  degrees of freedom
## AIC: 225.01
```

Backward Elimination



Backward Selection – SAS

```
proc logistic data=logistic.lowbwt plots(only)=(oddsratio);
   class race(ref='white') / param=ref;
   model low(event='1') = age race lwt smoke /
                          selection=backward slstay=0.03
                          clodds=pl clparm=pl;
   title 'Modeling Low Birth Weight';
run;
quit;
```

Backward Selection – SAS

Note: No (additional) effects met the 0.03 significance level for removal from the model.

Summary of Backward Elimination					
Step	**Effect Removed**	**DF**	**Number In**	**Wald Chi-Square**	**Pr > ChiSq**
1	**age**	1	3	0.4326	0.5107
2	**lwt**	1	2	4.4149	0.0356

Backward Selection – SAS

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
race	2	9.1128	0.0105
smoke	1	9.1357	0.0025

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.8405	0.3529	27.2065	<.0001
race	black	1	1.0841	0.4900	4.8951	0.0269
race	other	1	1.1086	0.4003	7.6689	0.0056
smoke		1	1.1160	0.3692	9.1357	0.0025

Backward Selection – R

```
back.model <- step(full.model, direction = "backward")
```

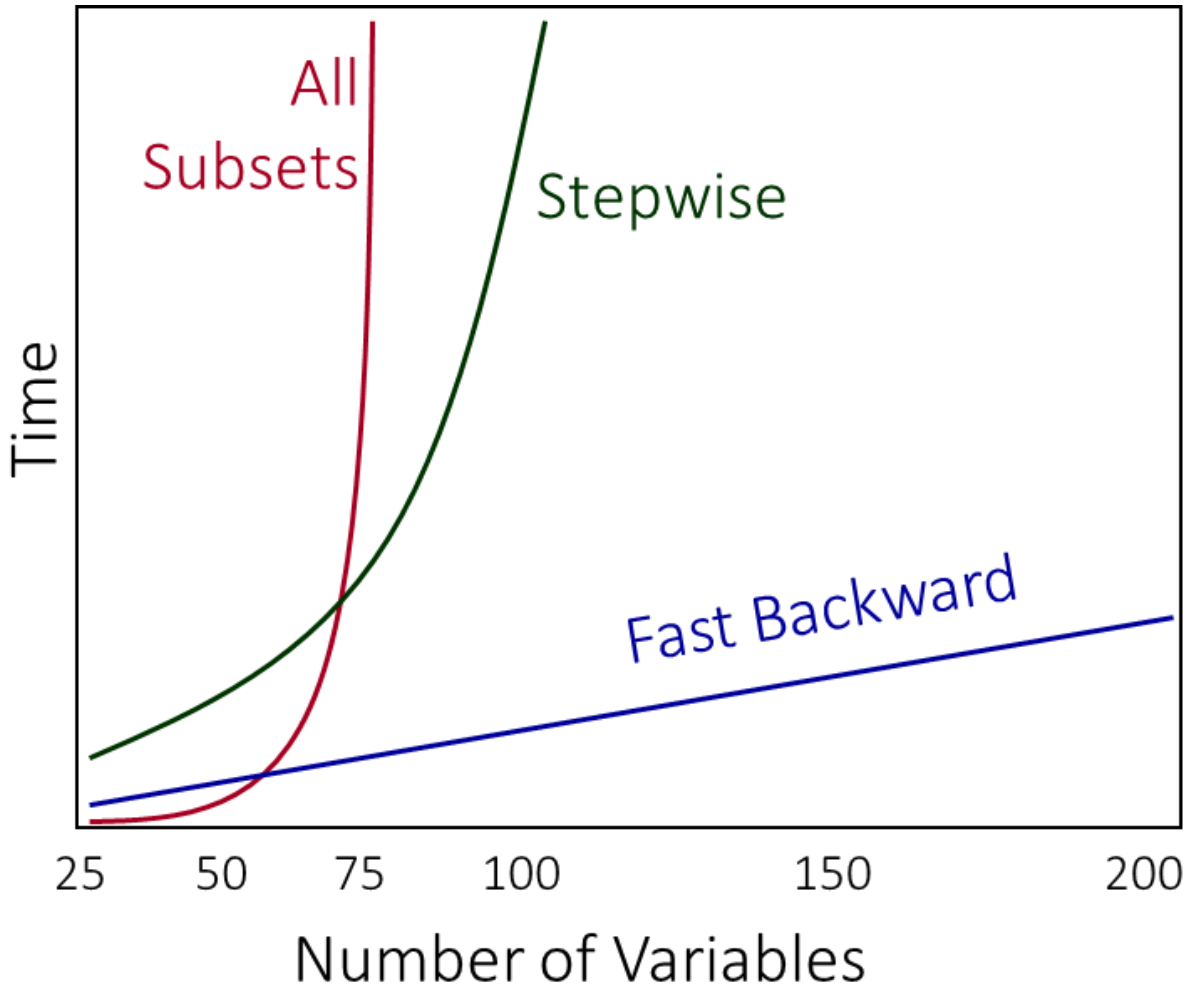
```
## Start:  AIC=226.58
## low ~ age + lwt + factor(smoke) + factor(race)
##
##                 Df Deviance    AIC
## - age            1   215.01 225.01
## <none>               214.58 226.58
## - lwt            1   218.86 228.86
## - factor(race)   2   222.88 230.88
## - factor(smoke)  1   222.66 232.66
##
## Step:  AIC=225.01
## low ~ lwt + factor(smoke) + factor(race)
##
##                 Df Deviance    AIC
## <none>               215.01 225.01
## - lwt            1   219.97 227.97
## - factor(race)   2   224.34 230.34
## - factor(smoke)  1   223.26 231.26
```

Backward Selection – R

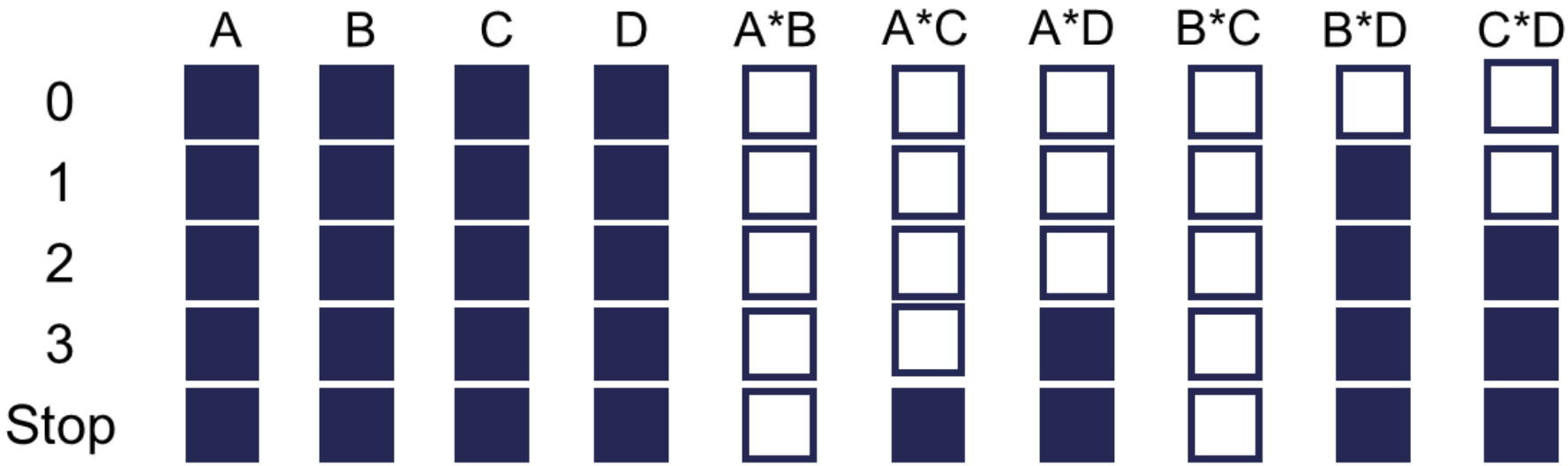
```r
summary(back.model)
```

```
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        1.18087    1.00983   1.169  0.24225
## lwt               -0.01326    0.00631  -2.101  0.03562 *
## factor(smoke)1     1.06001    0.37832   2.802  0.00508 **
## factor(race)other -0.31958    0.52560  -0.608  0.54317
## factor(race)white -1.29009    0.51087  -2.525  0.01156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 215.01  on 184  degrees of freedom
## AIC: 225.01
```

Scalability in PROC LOGISTIC



Interactions with Forward Selection



P-value vs. BIC Selection

- For our birth weight data set, BIC selection is the same as the p-value selection with the following alpha:

$$1 - P\left(\chi_1^2 > \log(n)\right) = 1 - P\left(\chi_1^2 > \log(189)\right) = 0.022$$

- Lot of attention being given to p-values and how other selection techniques are better.
- Attention **should** be on significance level (α), **not** on p-value.
- **DON'T ALWAYS USE 0.05!**

DIAGNOSTICS

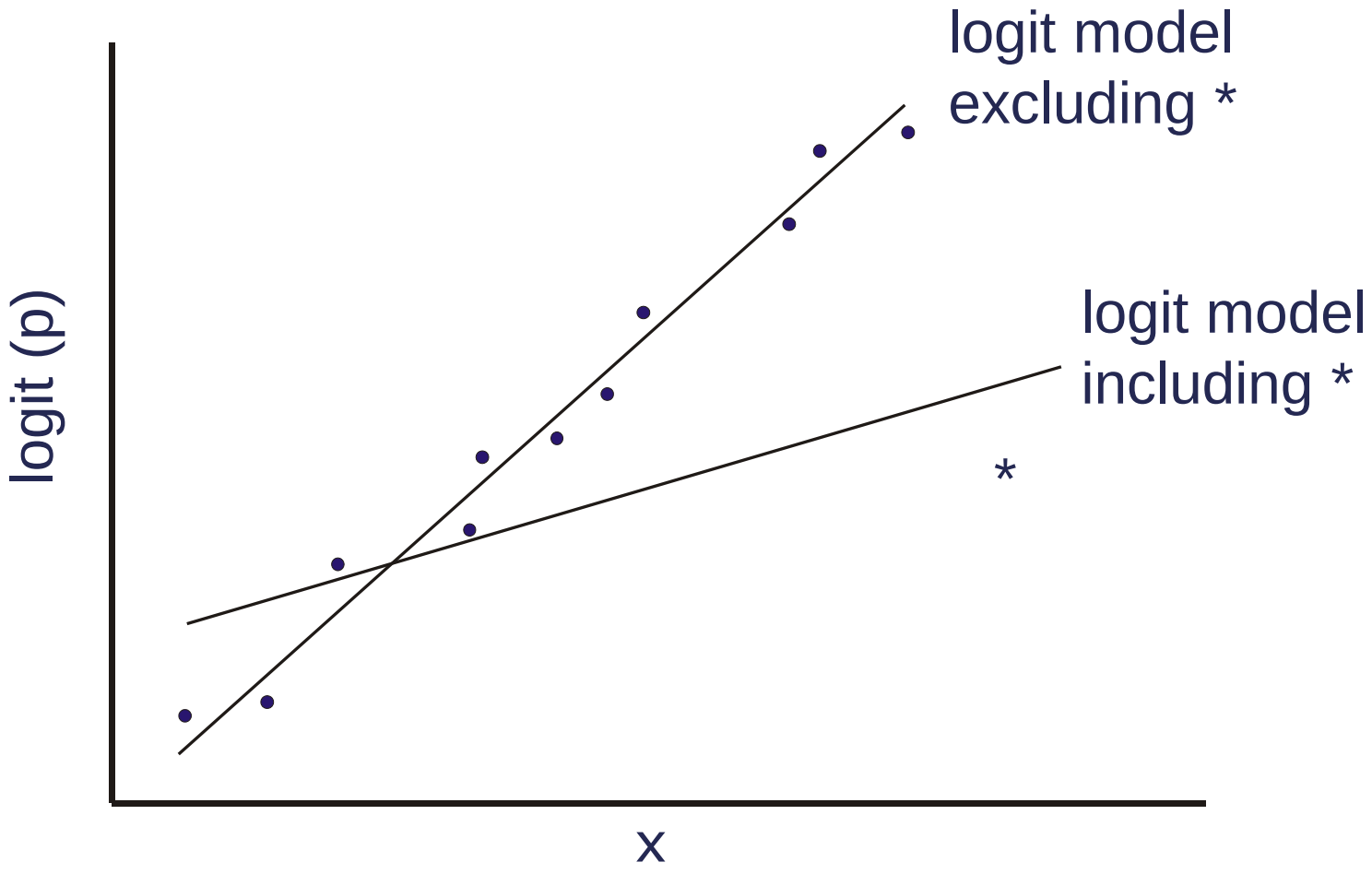
Residuals?

- Linear regression residuals have properties useful for model diagnostics.
- What is a residual in a binary response model?
- Many types of “residuals” in binary response model setting, just not as intuitive.
  - Deviance residuals
  - Partial residuals
  - Pearson residuals
  - Etc.

Deviance

- Model is a summary of a data set.
- The **saturated** model fits the data perfectly, but isn't really a useful summary.
- **Deviance** is a measure of how far a fitted model is from the saturated model – essentially our "error."
- Logistic regression minimizes the sum of squared deviances!
- Deviance residuals tell us how much each observation reduces the deviance.

Influence Statistics



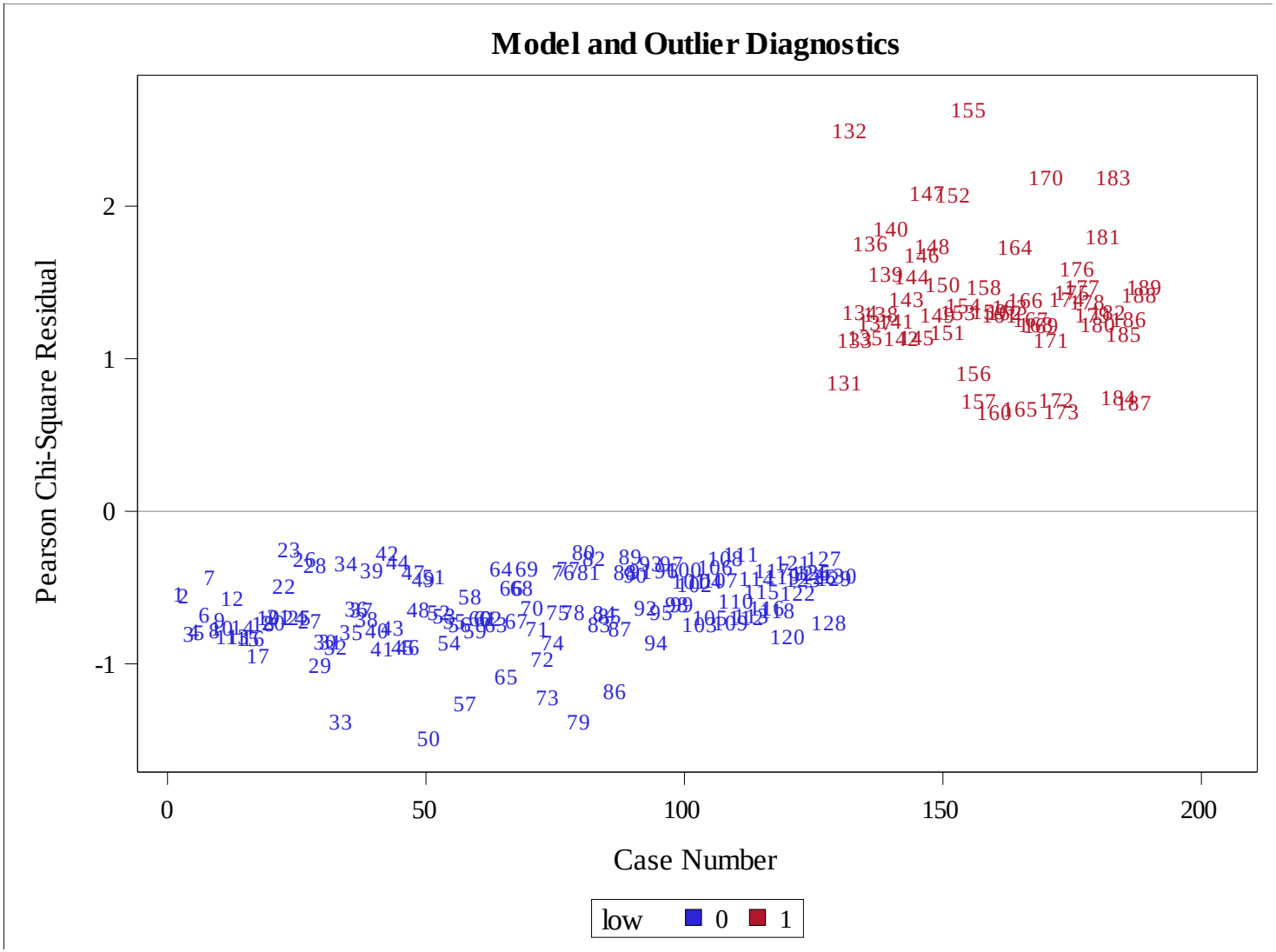
Influence Statistics

- DIFDEV
  - Measures change in deviance with deletion of the observation.
- DIFCHISQ
  - Measures change in Pearson Chi-square with deletion of observation.
- DFBETAS
  - Measure standardized change in each parameter estimate with deletion of observation.
- Cook’s D
  - Measures the overall impact to the coefficients in the model.

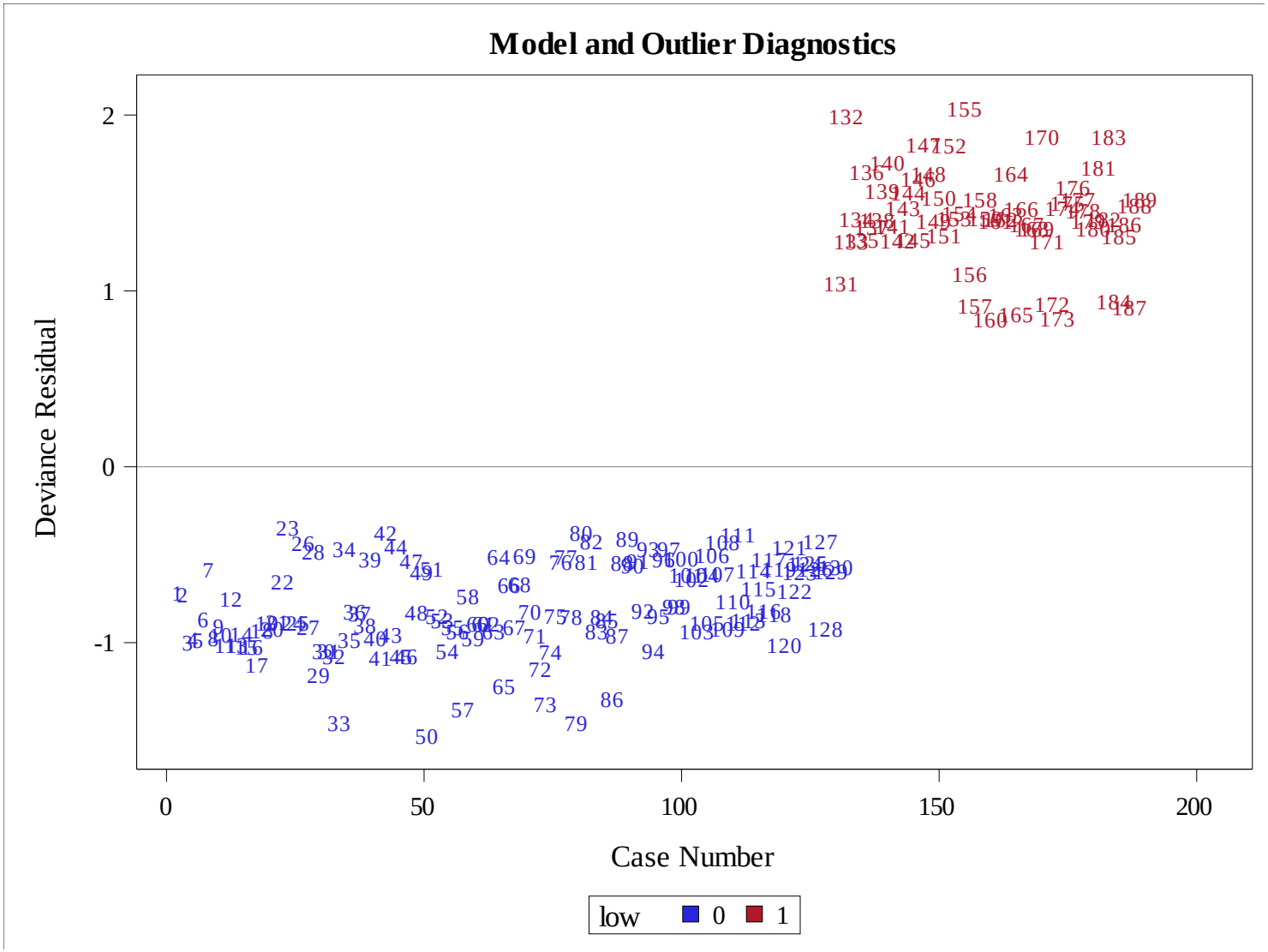
Diagnostic Plots – SAS

```
proc logistic data=logistic.lowbwt plots(unpack only label) =
                                    (influence dpc dfbetas);
   class race(ref='white') / param=ref;
   model low(event='1') = race lwt smoke;
   title 'Modeling Low Birth Weight';
run;
quit;
```

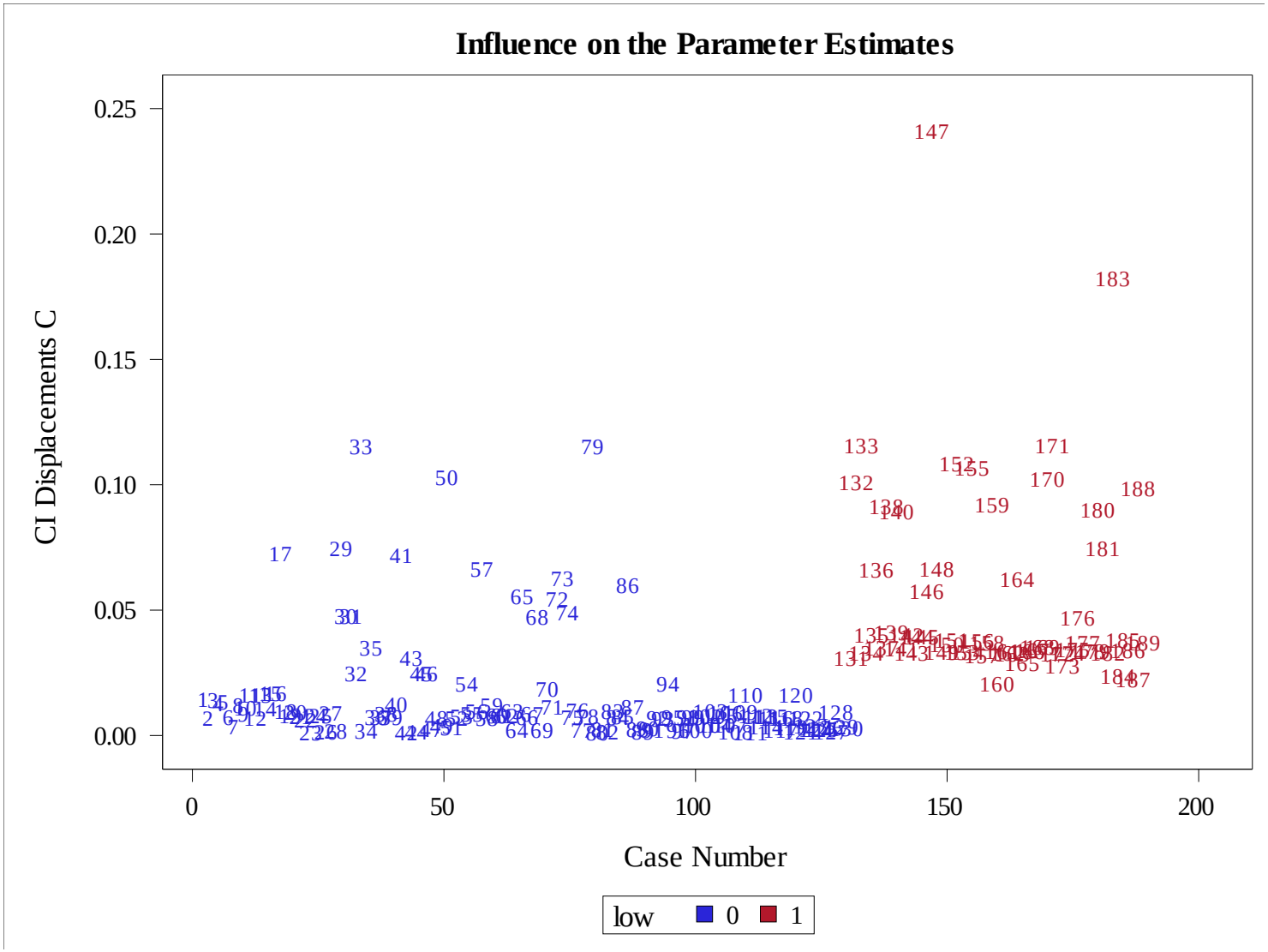
Diagnostic Plots – SAS



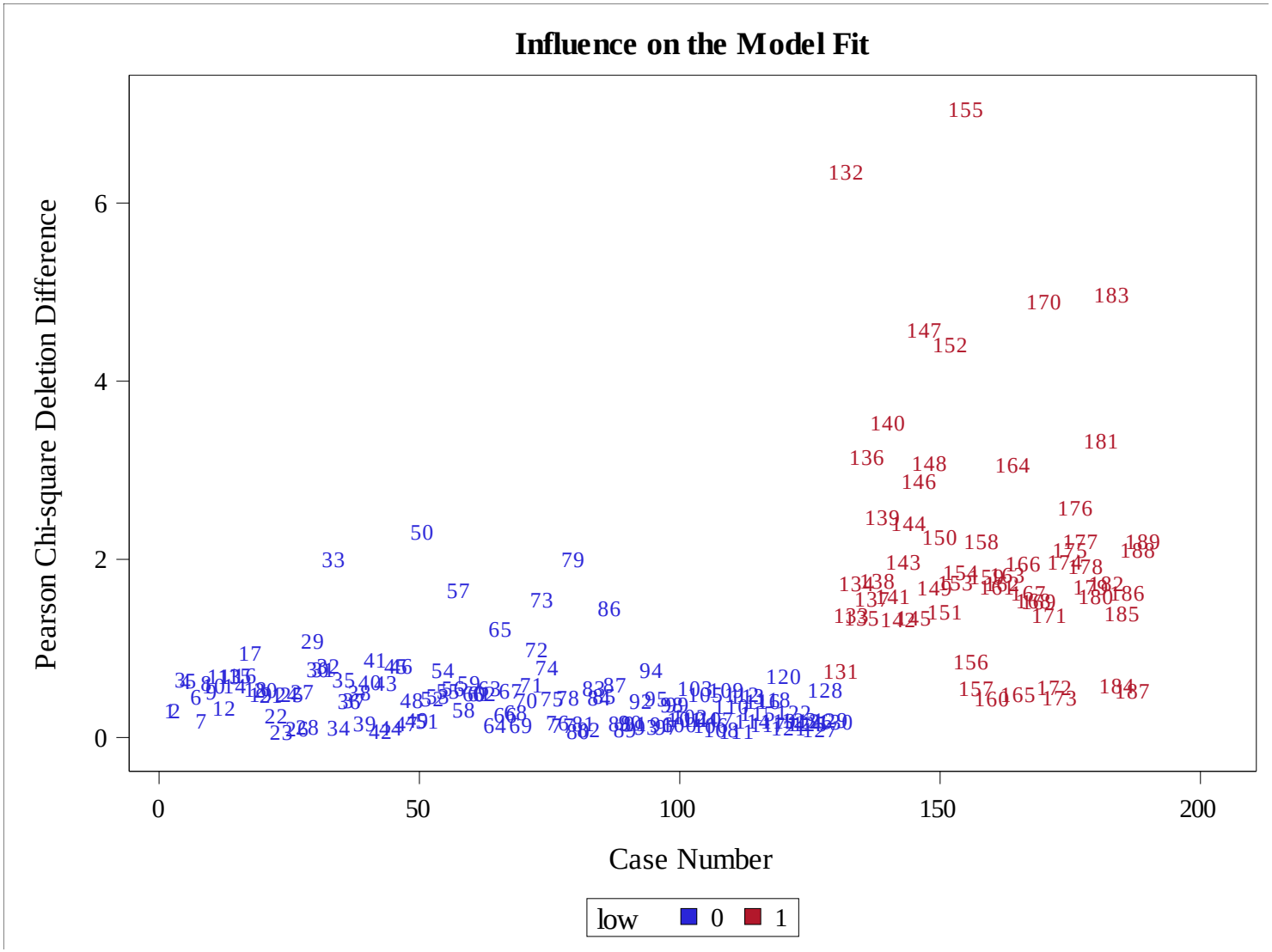
Diagnostic Plots – SAS



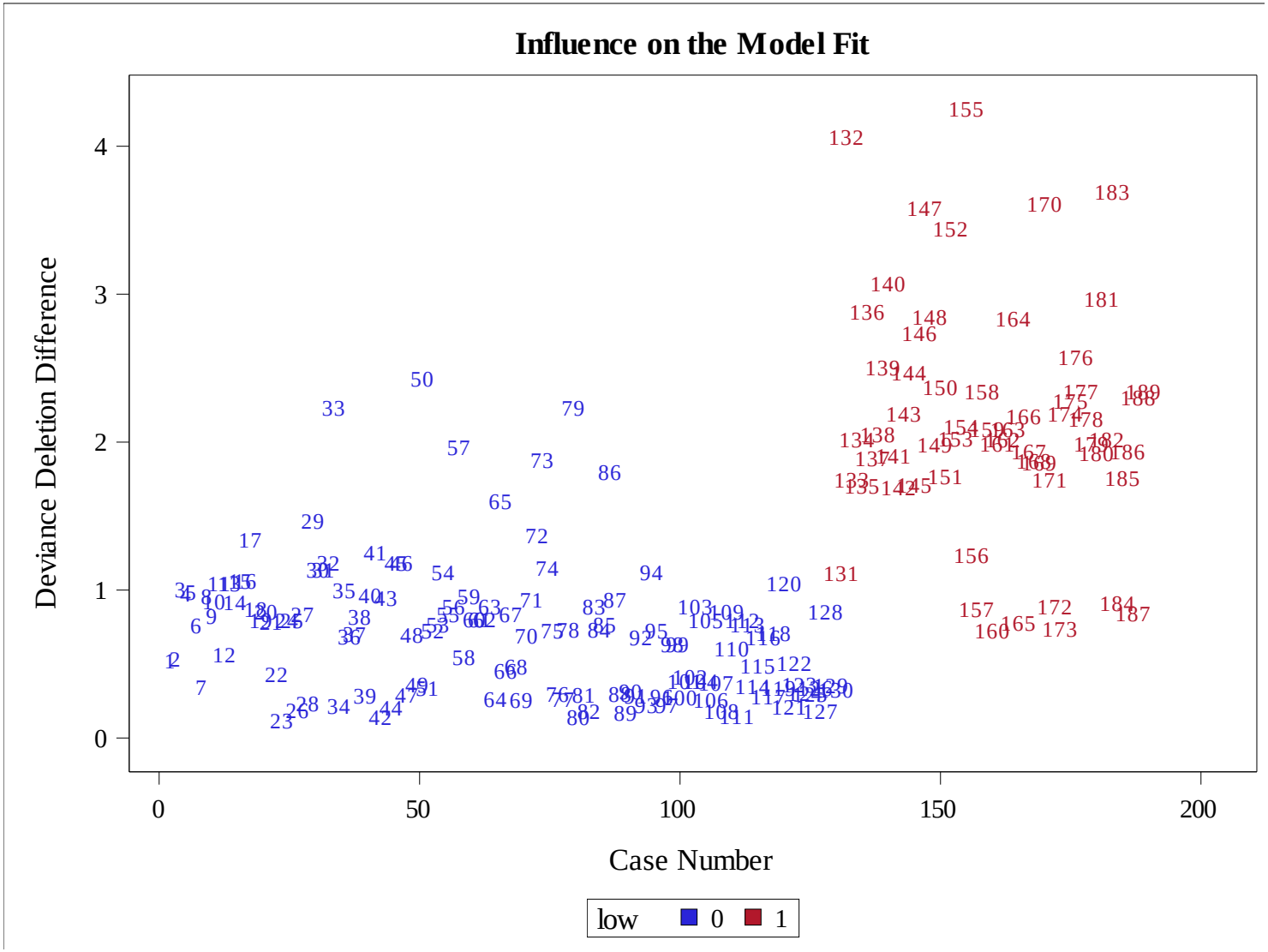
Diagnostic Plots – SAS



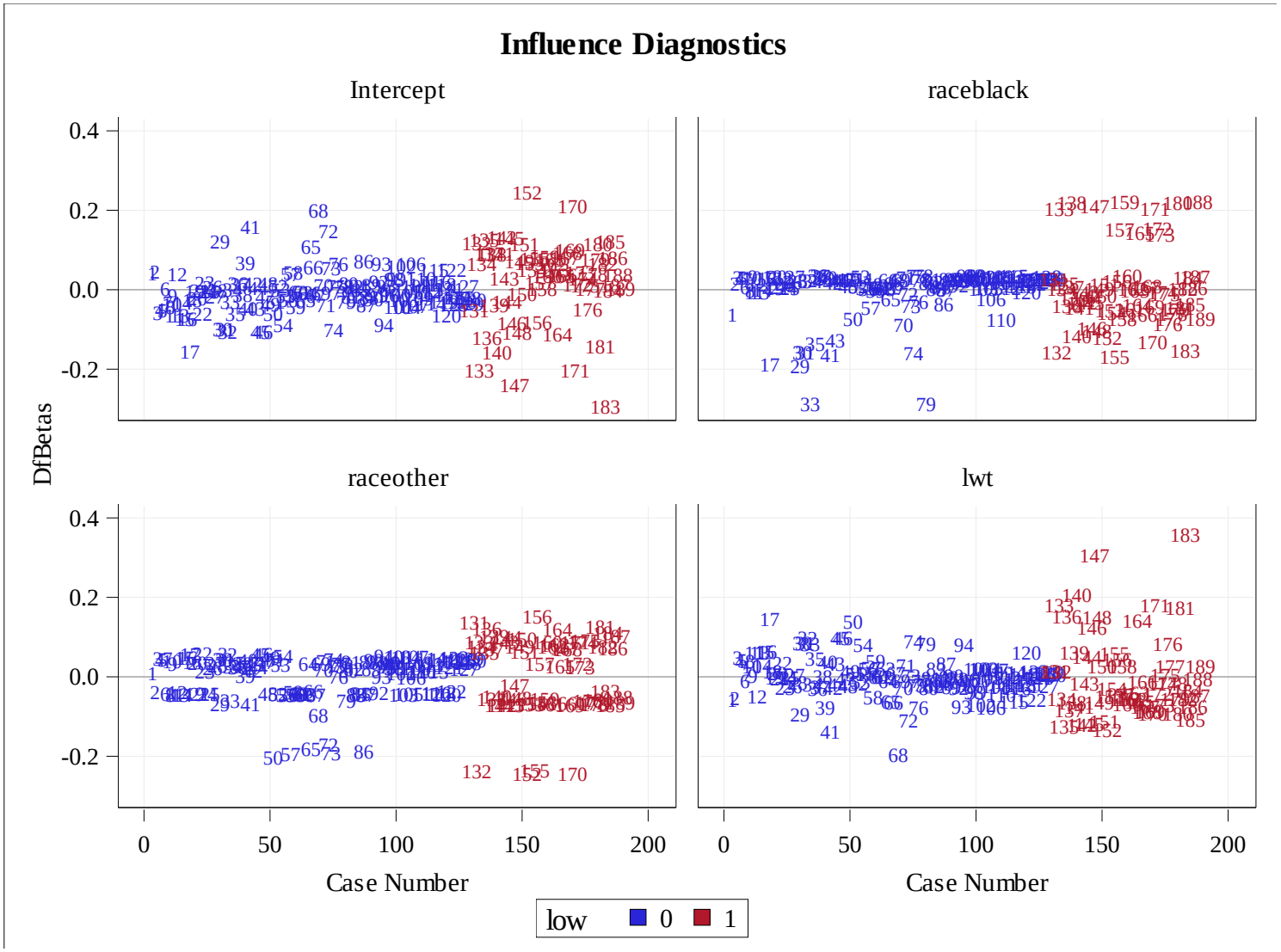
Diagnostic Plots – SAS



Diagnostic Plots – SAS



Diagnostic Plots – SAS



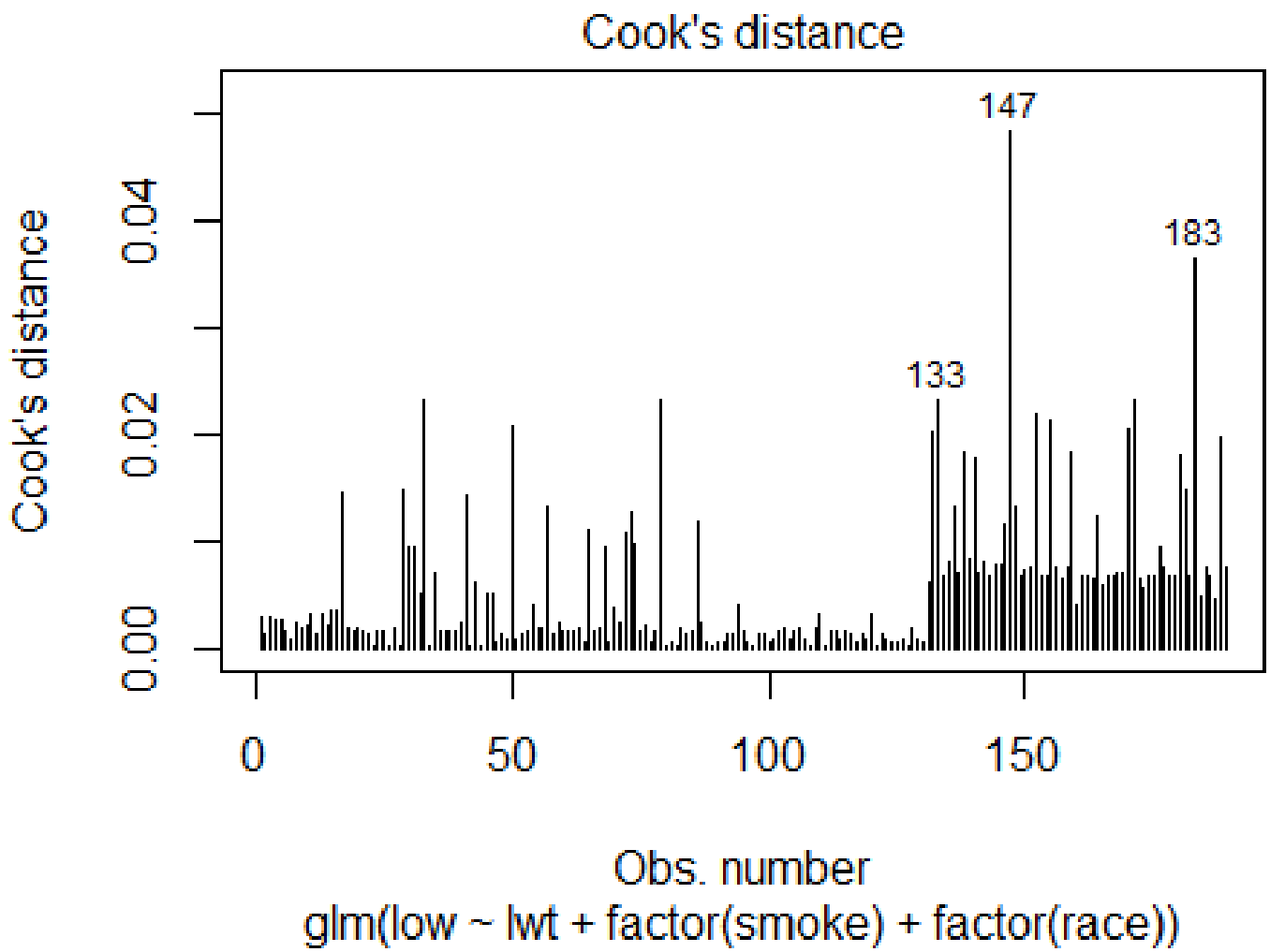
Diagnostic Plots – R

```
influence.measures(logit.model) # Prints for ALL obs. #

plot(logit.model, 4) # Cook's D Plot #

dfbetasPlots(logit.model, terms = "lwt", id.n = 5,
             col = ifelse(logit.model$y == 1, "red", "blue"))
```

Diagnostic Plots – R



Diagnostic Plots – R