

DIAGNOSTICS & SUBSET SELECTION

Dr. Aric LaBarr

Institute for Advanced Analytics

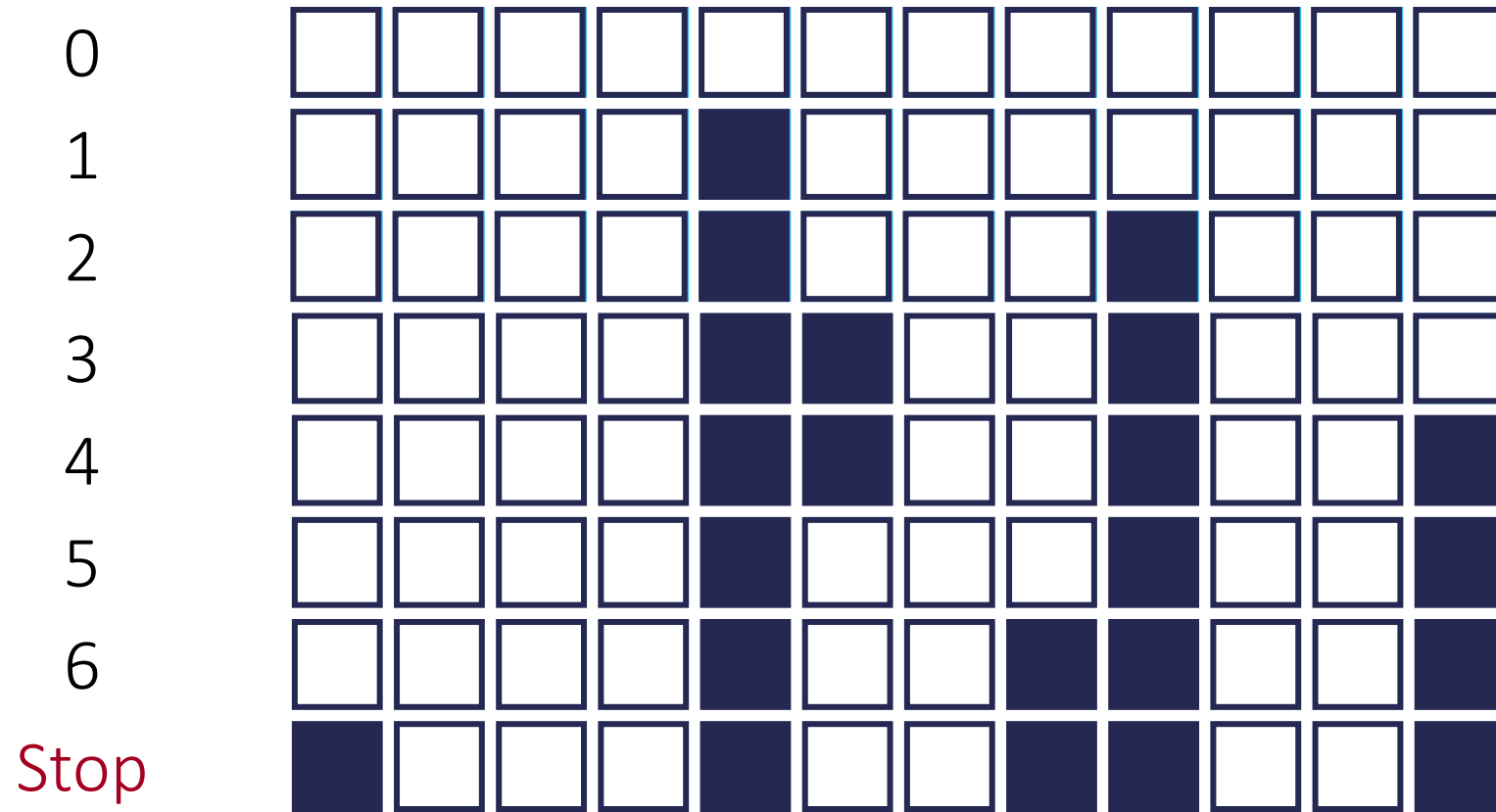
SUBSET SELECTION METHODS

Ames Real Estate Data

- 2930 homes in Ames, Iowa in the early 2000's.
- Physical attributes of homes along with sales price of home.



Stepwise Selection



Stepwise Selection

```
full.model <- glm(Bonus ~ Gr_Liv_Area + factor(House_Style) + Garage_Area +  
                  Fireplaces + factor(Full_Bath) + factor(Half_Bath) +  
                  Lot_Area + factor(Central_Air) + Second_Flr_SF +  
                  TotRms_AbvGrd + First_Flr_SF,  
                  data = train, family = binomial(link = "logit"))  
  
empty.model <- glm(Bonus ~ 1,  
                   data = train, family = binomial(link = "logit"))  
  
step.model <- step(empty.model,  
                   scope = list(lower=formula(empty.model),  
                                upper=formula(full.model)),  
                   direction = "both")
```

Stepwise Selection

Start: AIC=2777.81

Bonus ~ 1

	Df	Deviance	AIC
+ factor(Full_Bath)	4	1911.5	1921.5
+ Gr_Liv_Area	1	1926.4	1930.4
+ Garage_Area	1	2135.4	2139.4
+ First_Flr_SF	1	2294.1	2298.1
+ Fireplaces	1	2423.7	2427.7
+ TotRms_AbvGrd	1	2449.7	2453.7
+ factor(House_Style)	7	2542.1	2558.1
+ factor(Half_Bath)	2	2608.1	2614.1
+ Lot_Area	1	2621.9	2625.9
+ Second_Flr_SF	1	2631.8	2635.8
+ factor(Central_Air)	1	2654.3	2658.3
<none>		2775.8	2777.8

■
■
■

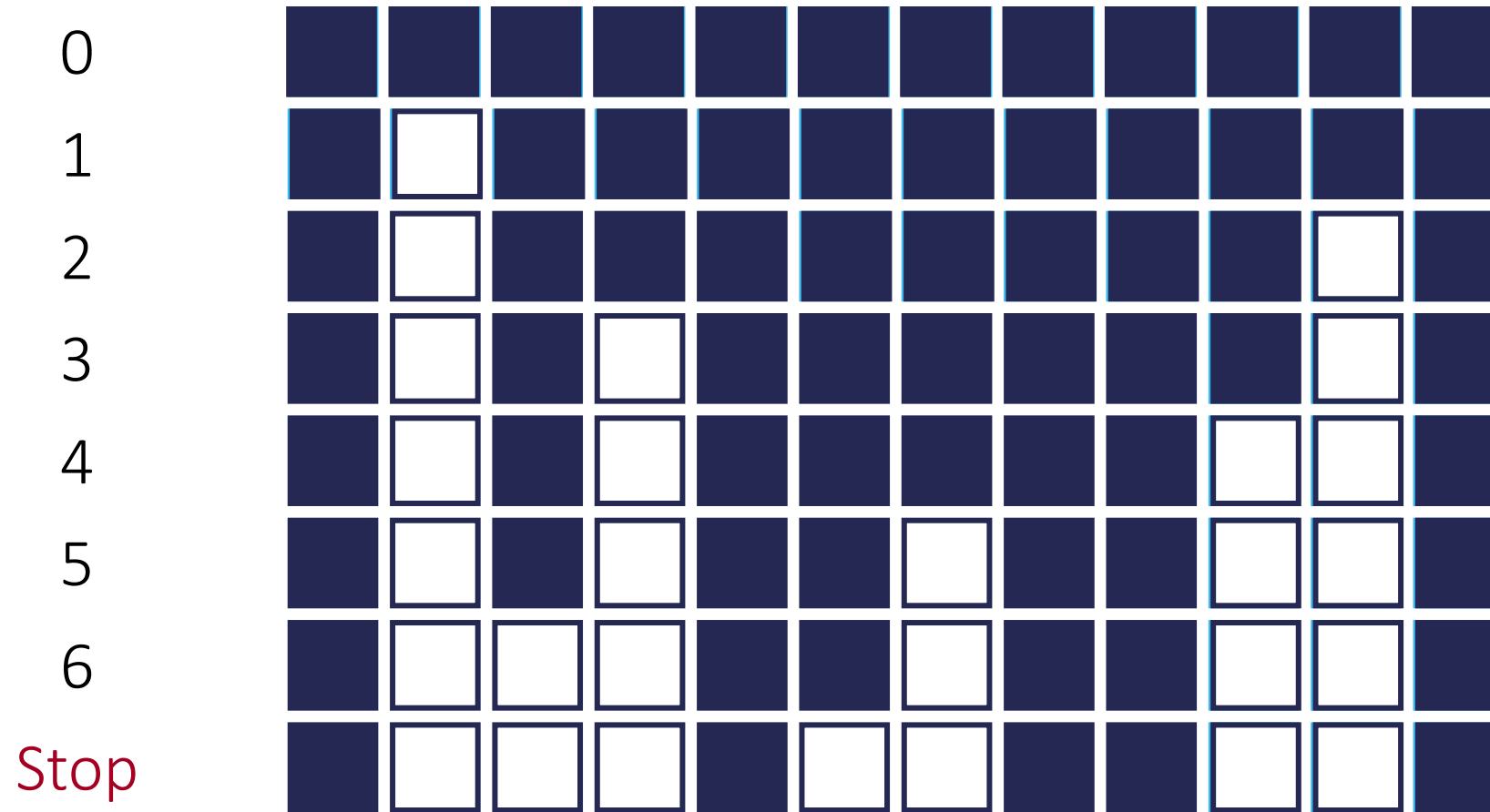
Stepwise Selection

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.041e+01	1.537e+00	-6.775	1.24e-11	***
factor(Full_Bath) 1	-6.860e-01	1.325e+00	-0.518	0.60469	
factor(Full_Bath) 2	1.894e+00	1.341e+00	1.412	0.15785	
factor(Full_Bath) 3	4.152e+00	1.610e+00	2.579	0.00991	**
factor(Full_Bath) 4	-1.261e+00	2.493e+00	-0.506	0.61305	
Garage_Area	3.583e-03	5.187e-04	6.907	4.96e-12	***
Fireplaces	9.142e-01	1.272e-01	7.186	6.67e-13	***
Gr_Liv_Area	3.827e-03	4.033e-04	9.488	< 2e-16	***
factor(House_Style)One_and_Half_Unf	-8.941e+00	3.682e+02	-0.024	0.98063	
factor(House_Style)One_Story	2.396e+00	3.285e-01	7.295	2.99e-13	***
factor(House_Style)SFoyer	1.760e+00	6.382e-01	2.757	0.00583	**
factor(House_Style)SLvl	1.105e+00	4.530e-01	2.438	0.01476	*
factor(House_Style)Two_and_Half_Fin	-4.855e-01	6.945e+00	-0.070	0.94427	
factor(House_Style)Two_and_Half_Unf	8.329e-01	8.891e-01	0.937	0.34890	
factor(House_Style)Two_Story	9.801e-01	3.380e-01	2.900	0.00373	**
factor(Half_Bath) 1	1.195e+00	2.153e-01	5.553	2.81e-08	***
factor(Half_Bath) 2	-1.301e-01	8.053e-01	-0.162	0.87163	
TotRms_AbvGrd	-4.322e-01	8.128e-02	-5.317	1.05e-07	***
factor(Central_Air)Y	1.620e+00	5.866e-01	2.762	0.00575	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Backward Elimination



Backward Selection

```
full.model <- glm(Bonus ~ Gr_Liv_Area + factor(House_Style) + Garage_Area +  
                  Fireplaces + factor(Full_Bath) + factor(Half_Bath) +  
                  Lot_Area + factor(Central_Air) + Second_Flr_SF +  
                  TotRms_AbvGrd + First_Flr_SF,  
                  data = train, family = binomial(link = "logit"))  
  
back.model <- step(full.model, direction = "backward")
```

Backward Selection

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.028e+01	1.541e+00	-6.673	2.51e-11	***
factor(House_Style)One_and_Half_Unf	-9.208e+00	3.686e+02	-0.025	0.98007	
factor(House_Style)One_Story	2.062e+00	4.945e-01	4.171	3.04e-05	***
factor(House_Style)SFoyer	1.464e+00	7.213e-01	2.030	0.04234	*
factor(House_Style)SLvl	9.390e-01	4.891e-01	1.920	0.05489	.
factor(House_Style)Two_and_Half_Fin	1.085e+00	6.908e+00	0.157	0.87524	
factor(House_Style)Two_and_Half_Unf	8.376e-01	8.904e-01	0.941	0.34687	
factor(House_Style)Two_Story	1.010e+00	3.498e-01	2.886	0.00390	**
Garage_Area	3.499e-03	5.210e-04	6.716	1.87e-11	***
Fireplaces	8.965e-01	1.279e-01	7.010	2.39e-12	***
factor(Full_Bath)1	-6.540e-01	1.330e+00	-0.492	0.62302	
factor(Full_Bath)2	1.930e+00	1.347e+00	1.433	0.15196	
factor(Full_Bath)3	4.355e+00	1.618e+00	2.691	0.00712	**
factor(Full_Bath)4	-1.073e+00	2.436e+00	-0.440	0.65971	
factor(Half_Bath)1	1.228e+00	2.215e-01	5.545	2.94e-08	***
factor(Half_Bath)2	-6.069e-02	8.103e-01	-0.075	0.94030	
factor(Central_Air)Y	1.590e+00	5.909e-01	2.690	0.00715	**
Second_Flr_SF	3.466e-03	6.632e-04	5.226	1.73e-07	***
TotRms_AbvGrd	-4.339e-01	8.142e-02	-5.329	9.86e-08	***
First_Flr_SF	4.011e-03	4.351e-04	9.220	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Comparison of Backward with Stepwise

Stepwise Selection Variables

- Full Bath
- Garage Area
- Fireplaces
- Greater Living Area
- House Style
- Half Bath
- Total Rooms (Above Ground)
- Central Air

Backward Selection Variables

- Full Bath
- Garage Area
- Fireplaces
- House Style
- Half Bath
- Total Rooms (Above Ground)
- Central Air
- 1st Floor Sqft
- 2nd Floor Sqft

Interactions with Forward Selection

	A	B	C	D	A*B	A*C	A*D	B*C	B*D	C*D
0	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
2	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
3	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Stop	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>



P-VALUE VS. AIC/BIC METRICS

P-value vs. AIC Selection

- Lot of attention being given to p-values and how other selection techniques are better.
- Some of these are **the same as** p-values...

$$AIC = -2 \log(L) + 2p$$

P-value vs. AIC Selection

- Lot of attention being given to p-values and how other selection techniques are better.
- Some of these are **the same as** p-values...

$$AIC = -2 \log(L) + 2p$$

Model better with
lower AIC...

$$AIC_{p+1} < AIC_p$$

P-value vs. AIC Selection

- Lot of attention being given to p-values and how other selection techniques are better.
- Some of these are **the same as** p-values...

$$AIC = -2 \log(L) + 2p$$

Model better with
lower AIC...

$$AIC_{p+1} < AIC_p$$

$$-2 \log(L_{p+1}) + 2(p + 1) < -2 \log(L_p) + 2(p)$$

$$2 < 2(\log(L_{p+1}) - \log(L_p))$$

P-value vs. AIC Selection

- Lot of attention being given to p-values and how other selection techniques are better.
- Some of these are **the same as** p-values...

$$AIC = -2 \log(L) + 2p$$

Model better with
lower AIC...

$$AIC_{p+1} < AIC_p$$

$$-2 \log(L_{p+1}) + 2(p + 1) < -2 \log(L_p) + 2(p)$$

$$2 < 2(\log(L_{p+1}) - \log(L_p)) \quad \text{LRT!}$$

P-value vs. AIC Selection

- Lot of attention being given to p-values and how other selection techniques are better.
- Some of these are **the same as** p-values...

$$AIC = -2 \log(L) + 2p$$

Model better with
lower AIC...

$$AIC_{p+1} < AIC_p$$

$$-2 \log(L_{p+1}) + 2(p+1) < -2 \log(L_p) + 2(p)$$

$$2 < \chi_1^2$$

P-value vs. AIC Selection

- Lot of attention being given to p-values and how other selection techniques are better.
- Some of these are **the same as** p-values...

$$AIC = -2 \log(L) + 2p$$

Model better with
lower AIC...

$$AIC_{p+1} < AIC_p$$

$$-2 \log(L_{p+1}) + 2(p+1) < -2 \log(L_p) + 2(p)$$

$$2 < \chi_1^2$$

Model better with
variable p-value
below sig. level...

$$1 - P(\chi_1^2 > 2) = 0.1573$$

P-value vs. BIC Selection

- Lot of attention being given to p-values and how other selection techniques are better.
- Some of these are **the same as** p-values...

$$BIC = -2 \log(L) + p \times \log(n)$$

Model better with
lower BIC...

$$BIC_{p+1} < BIC_p$$

$$-2 \log(L_{p+1}) + \log(n) (p + 1) < -2 \log(L_p) + \log(n) (p)$$

$$\log(n) < \chi_1^2$$

Model better with
variable p-value
below sig. level...

$$1 - P(\chi_1^2 > \log(n)) = \dots$$

P-value vs. BIC Selection

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

$$1 - P(\chi_1^2 > \log(n)) = 1 - P(\chi_1^2 > \log(2051)) = 0.0057$$

- 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!**



GOODNESS-OF-FIT

OPTIONAL: SELF-PACED STUDY

Calibration

- **Calibration** measures how well predicted probabilities agree with actual frequency counts of outcomes.
- Helps detect bias!
 - Are predictions systematically too low or too high?

Calibration Curve

- Curve **above** 45° line indicates the model is predicting **lower** probabilities than actually observed.
- Curve **below** 45° line indicates the model is predicting **higher** probabilities than actually observed.
- Caveat:
 - Calibration depends on the observed proportion of events in the data, so models will likely have poor calibration on out-of-sample data.
 - Best used for goodness-of-fit in training, not on validation.

Calibration Curve

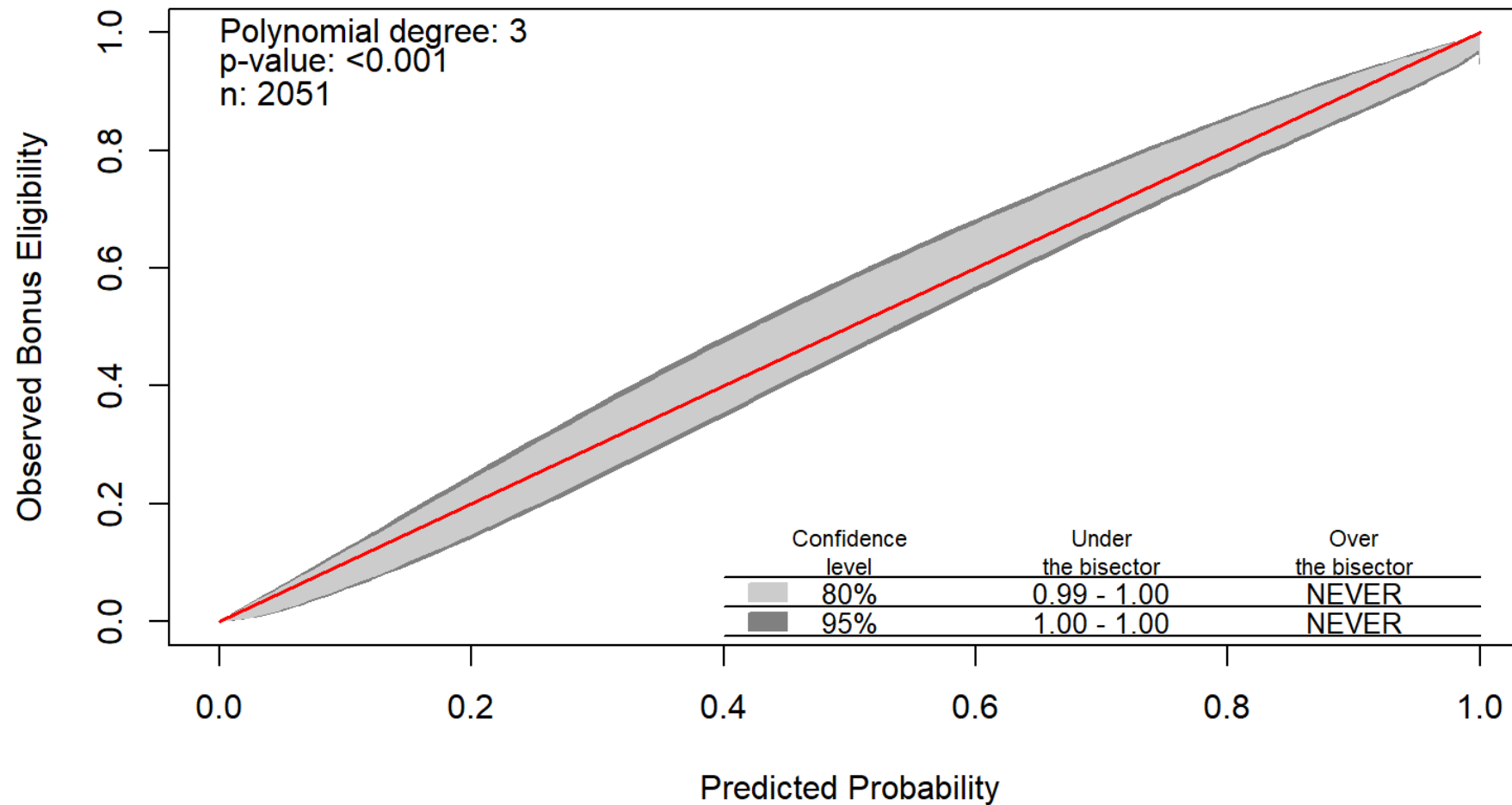
```
library(givitiR)

cali.curve <- givitiCalibrationBelt(o = train$Bonus,
                                   e = predict(logit.model, type = "response"),
                                   devel = "internal", maxDeg = 5)

plot(cali.curve, main = "Bonus Eligibility Model Calibration Curve",
     xlab = "Predicted Probability",
     ylab = "Observed Bonus Eligibility")
```

Calibration Curve

Bonus Eligibility Model Calibration Curve



DIAGNOSTICS

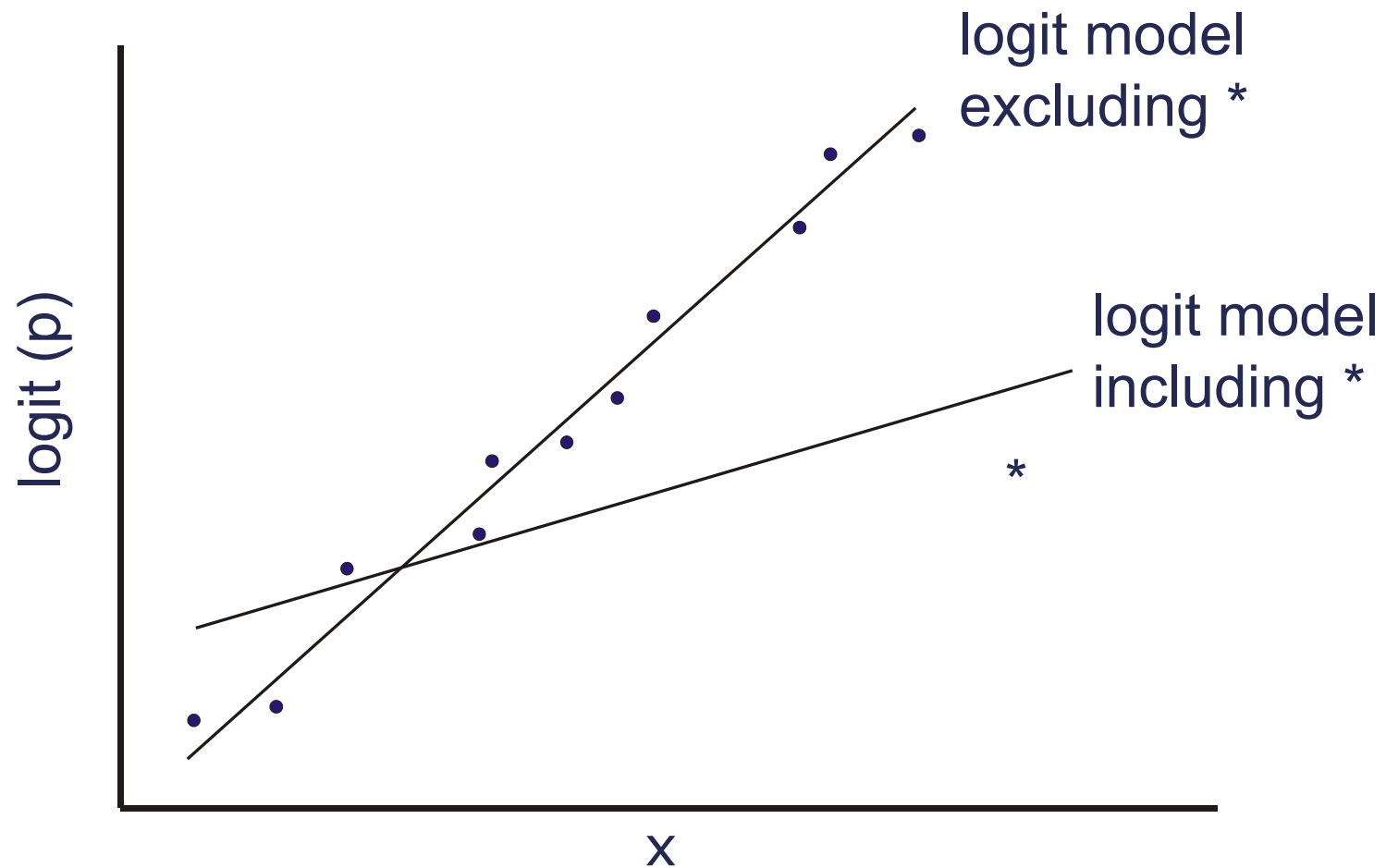
Residuals?

- 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.

Diagnostics

```
library(car)
```

```
logit.model <- glm(Bonus ~ Gr_Liv_Area + factor(House_Style) + Garage_Area +  
                  Fireplaces + factor(Full_Bath) + Lot_Area +  
                  factor(Central_Air) + TotRms_AbvGrd +  
                  Gr_Liv_Area:Fireplaces,  
                  data = train, family = binomial(link = "logit"))
```

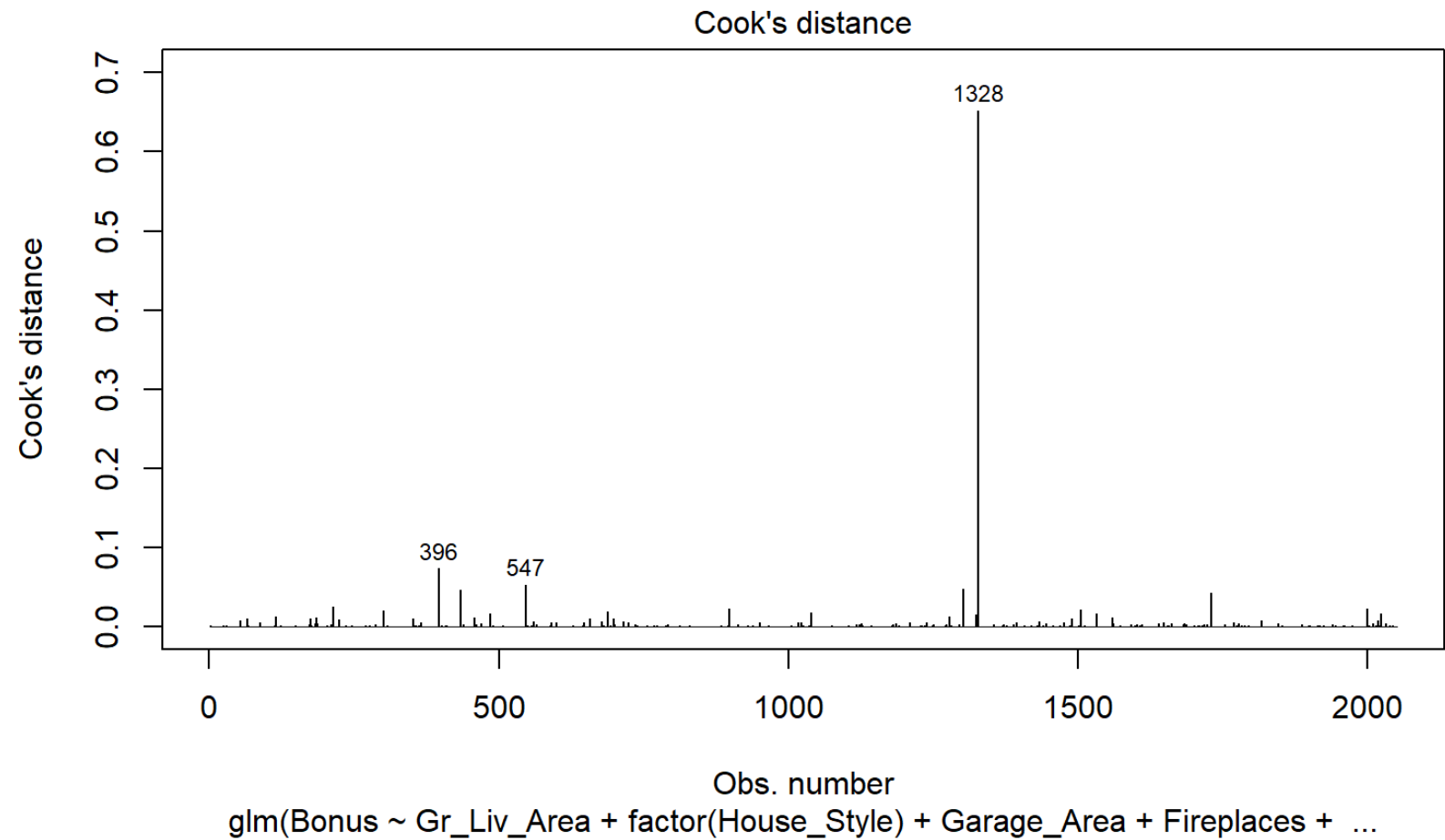
```
influence.measures(logit.model)$infmt
```



Prints out metrics previously listed
for every observation!
Output not shown here.

Diagnostics

```
plot(logit.model, 4)
```



Diagnostics

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

