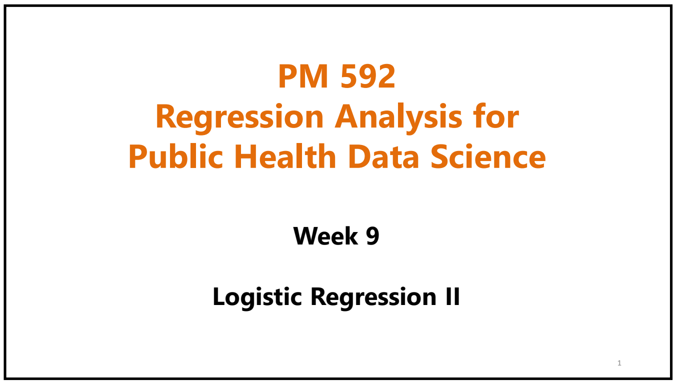


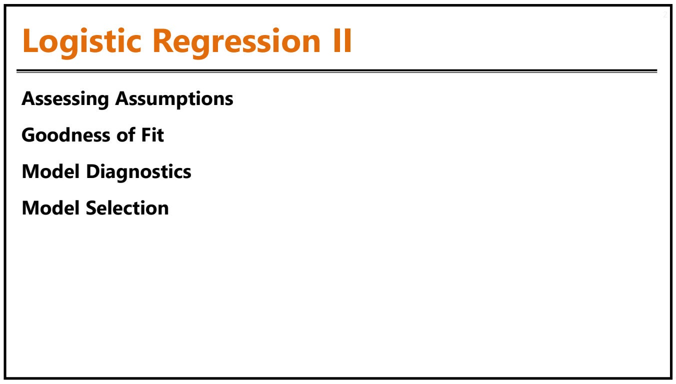
PM 592
Regression Analysis for Public Health Data Science

Week 9

Logistic Regression II

1

1



Logistic Regression II

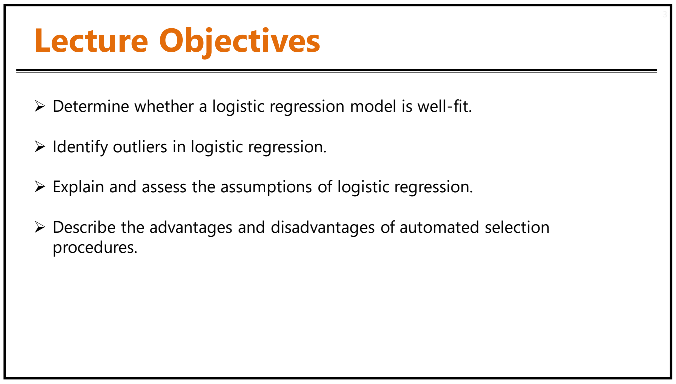
Assessing Assumptions

Goodness of Fit

Model Diagnostics

Model Selection

2



Lecture Objectives

- Determine whether a logistic regression model is well-fit.
- Identify outliers in logistic regression.
- Explain and assess the assumptions of logistic regression.
- Describe the advantages and disadvantages of automated selection procedures.

3

1. Review

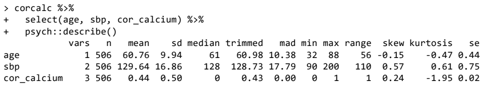
- ✓ Three ways to measure the effect on a binary outcome
- ✓ 2x2 contingency tables, odds, the odds ratio
- ✓ The concept of a "link" function
- ✓ The logit link – computing an odds ratio
- ✓ The logit link – computing predicted probabilities

2. Assessing Assumptions

Example

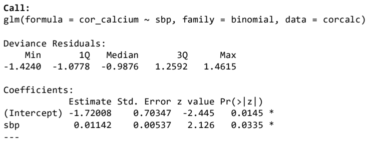
In a study of 508 adults, vital characteristics (e.g. blood pressure, height, weight) and presence of coronary calcium (a measure of blockage in the arteries of the heart) was assessed.

What is the relationship between age and SBP with presence of coronary calcium?



```
> corcalc %>%
+   select(age, sbp, cor_calcium) %>%
+   psych::describe()
            vars   n   mean    sd median trimmed   mad min max range  skew kurtosis   se
age            1 506  60.76  9.94     61   60.98 10.38  32  88    56 -0.15    -0.47 0.44
sbp            2 506 129.64 16.86    128  128.73 17.79  90 200   110  0.57     0.61 0.75
cor_calcium    3 506   0.44  0.50      0    0.43  0.00   0   1     1  0.24    -1.95 0.02
```

2. Assessing Assumptions



```
Call:
glm(formula = cor_calcium ~ sbp, family = binomial, data = corcalc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max  
-1.4240  -1.0778  -0.9876   1.2592   1.4615  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)  
(Intercept) -1.72008    0.70347  -2.445   0.0145 *
sbp          0.01142    0.00537   2.126   0.0335 *
---
```

SBP is significantly related to coronary calcium. The odds ratio associated with a 1-unit increase in SBP is exp(0.01142) = 1.011. Since this is a small odds ratio, it might help to instead interpret the odds ratio for a 10-unit increase in SBP. This odds ratio would be exp(10* 0.01142) = 1.12.

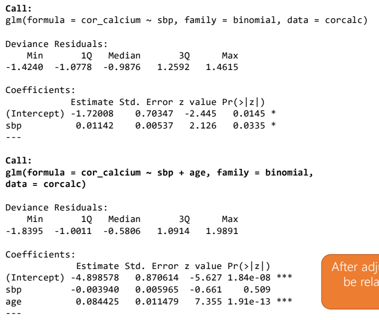
"A 10-unit increase in SBP is associated with 1.12 times the odds of coronary calcium."

"A 10-unit increase in SBP increases the likelihood of coronary calcium by 12%."

We know this is a logistic regression because we specified "family = binomial".

2. Assessing Assumptions

7



```
Call:
glm(formula = cor_calcium ~ sbp, family = binomial, data = corcalc)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.4240  -1.0778  -0.9876   1.2592   1.4615  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)  
(Intercept) -1.72008    0.70347  -2.445   0.0145 *
sbp          0.01142    0.00537   2.126   0.0335 *
---

Call:
glm(formula = cor_calcium ~ sbp + age, family = binomial, 
data = corcalc)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.8395  -1.0011  -0.5806   1.0914   1.9891  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) -4.898578   0.870614  -5.627 1.84e-08 ***
sbp         -0.003940   0.005965  -0.661    0.509    
age          0.084425   0.011479   7.355 1.91e-13 ***
---
```

After adjusting for age, SBP does not appear to be related to presence of coronary calcium (p=.51).

7

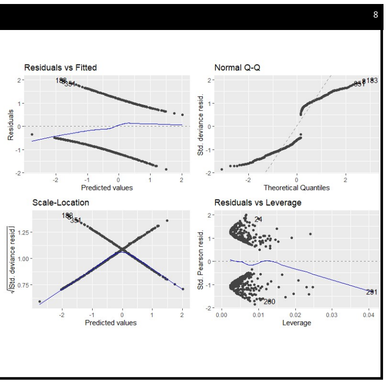
2. Assessing Assumptions

8

How do the residuals look?

Because we are comparing observed values of Y (that can only take on the values of 0 and 1) with predicted probabilities $\hat{\pi}$, our residuals are going to look a lot weirder than usual.

In fact, the assumptions of OLS (ordinary least squares) regression do not apply for this type of modeling.



8

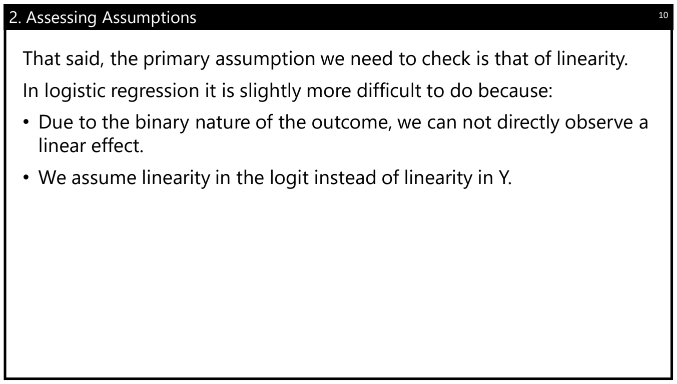
2. Assessing Assumptions

9

Here, we will go over the usual assumptions of linear regression and see how they apply to logistic regression.

- **Linearity** – X and Y cannot be linearly related if Y is binary. However we do assume linearity *in the logit*.
- **Independence** – we do assume all X are independent of each other.
- **Normality** – we do not assume that the residuals are normally distributed.
- **Equal Variances** – we do not assume that the residuals have constant variance over all X values.

9



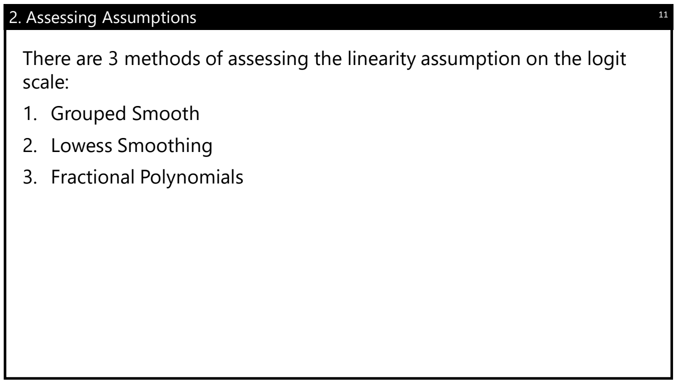
2. Assessing Assumptions

That said, the primary assumption we need to check is that of linearity.

In logistic regression it is slightly more difficult to do because:

- Due to the binary nature of the outcome, we can not directly observe a linear effect.
- We assume linearity in the logit instead of linearity in Y.

10

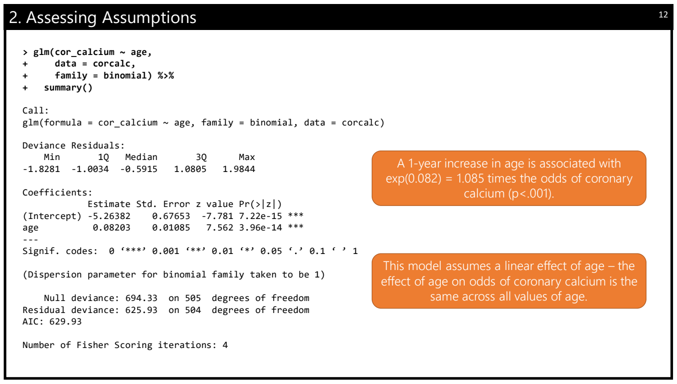


2. Assessing Assumptions

There are 3 methods of assessing the linearity assumption on the logit scale:

1. Grouped Smooth
2. Lowess Smoothing
3. Fractional Polynomials

11



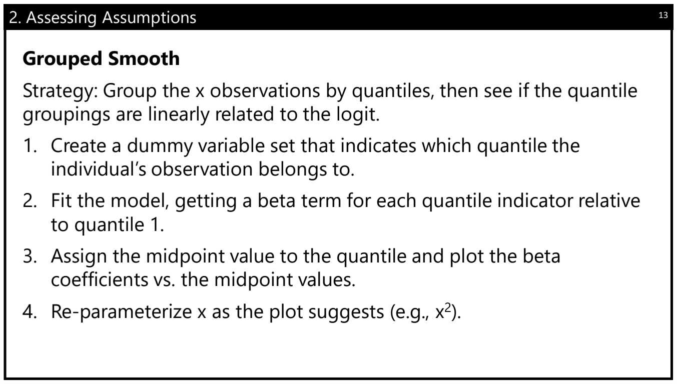
2. Assessing Assumptions

```
> glm(cor_calcium ~ age,
+     data = corcalc,
+     family = binomial) %>%
+   summary()

Call:
glm(formula = cor_calcium ~ age, family = binomial, data = corcalc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8281  -1.0034  -0.5915   1.0805   1.9844

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.26382    0.67653  -7.781 7.22e-15 ***
age          0.08203    0.01085   7.562 3.96e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 694.33  on 505  degrees of freedom
Residual deviance: 625.93  on 504  degrees of freedom
AIC: 629.93

Number of Fisher Scoring iterations: 4
```

A 1-year increase in age is associated with exp(0.082) = 1.085 times the odds of coronary calcium (p<.001).

This model assumes a linear effect of age – the effect of age on odds of coronary calcium is the same across all values of age.

12



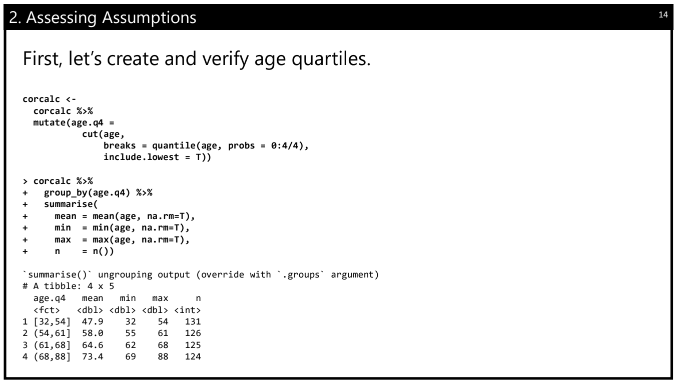
2. Assessing Assumptions 13

Grouped Smooth

Strategy: Group the x observations by quantiles, then see if the quantile groupings are linearly related to the logit.

1. Create a dummy variable set that indicates which quantile the individual's observation belongs to.
2. Fit the model, getting a beta term for each quantile indicator relative to quantile 1.
3. Assign the midpoint value to the quantile and plot the beta coefficients vs. the midpoint values.
4. Re-parameterize x as the plot suggests (e.g., x^2).

13

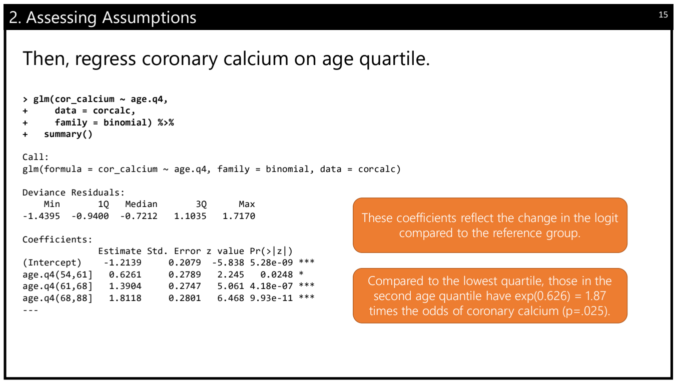


2. Assessing Assumptions 14

First, let's create and verify age quartiles.

```
corcalc <-
  corcalc %>%
  mutate(age.q4 =
           cut(age,
               breaks = quantile(age, probs = 0:4/4),
               include.lowest = T))

> corcalc %>%
+   group_by(age.q4) %>%
+   summarise(
+     mean = mean(age, na.rm=T),
+     min  = min(age, na.rm=T),
+     max  = max(age, na.rm=T),
+     n    = n())

`summarise()` ungrouping output (override with `.groups` argument)
# A tibble: 4 x 5
  age.q4   mean   min   max     n
  <fct>   <dbl> <dbl> <dbl> <int>
1 [32,54]  47.9    32    54   131
2 (54,61]  58.0    55    61   126
3 (61,68]  64.6    62    68   125
4 (68,88]  73.4    69    88   124
```

14



2. Assessing Assumptions 15

Then, regress coronary calcium on age quartile.

```
> glm(cor_calcium ~ age.q4,
+     data = corcalc,
+     family = binomial) %>%
+   summary()

Call:
glm(formula = cor_calcium ~ age.q4, family = binomial, data = corcalc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4395  -0.9400  -0.7212   1.1035   1.7170

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.2139     0.2079  -5.838 5.28e-09 ***
age.q4(54,61]   0.6261     0.2789   2.245   0.0248 *
age.q4(61,68]   1.3904     0.2747   5.061 4.18e-07 ***
age.q4(68,88]   1.8118     0.2801   6.468 9.93e-11 ***
---
```

These coefficients reflect the change in the logit compared to the reference group.

Compared to the lowest quartile, those in the second age quantile have exp(0.626) = 1.87 times the odds of coronary calcium (p=.025).

15

2. Assessing Assumptions

The global test (Likelihood Ratio vs. the null model) shows us that these variables, as a set, are related to coronary calcium.

```
> glm(cor_calcium ~ age.q4,
+     data = corcalc,
+     family = binomial) %>%
+   anova(test = “LRT")
Analysis of Deviance Table

Model: binomial, link: logit

Response: cor_calcium

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     505     694.33
age.q4  3   55.501       502     638.83 5.368e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1---
```

16

2. Assessing Assumptions

When we use dummy predictor variables, we allow for modeling **flexibility** because we don’t assume a linear relationship across all X values.

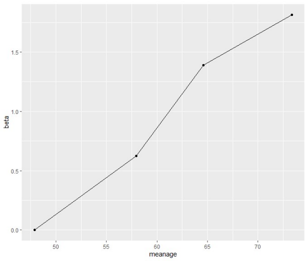
If we plot the logit and see that the relationship between X and the logit appears linear, then we know we can be more restrictive in our modeling approach.

17

2. Assessing Assumptions

The relationship between the logit and age quartile isn’t perfectly linear, but it seems like a pretty good approximation!

In this approach, we allow flexibility in the estimation of the logit among age quartiles.



18

2. Assessing Assumptions

The logit is estimated as a function of the dummy variables for age quartile.

$$logit(\hat{\pi}) = \beta_0 + \beta_1 X_{age.q2} + \beta_2 X_{age.q3} + \beta_3 X_{age.q4}$$

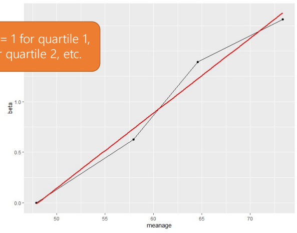
If a linear approach is good enough, though, then we could fit this relationship with a straight line.

2. Assessing Assumptions

The equation for the red line is simpler but imposes more constraints.

$$logit(\hat{\pi}) = \beta_0 + \beta_1 X_{ageq}$$

The dummy variable scheme is more flexible in comparison to the linear model. We can use the likelihood ratio test to see if this flexibility improves model fit, or if we should stay with the more parsimonious linear model.



ageq = 1 for quartile 1, 2 for quartile 2, etc.

```
> anova(agequantlin.m, agequant.m, test = "LRT")
Analysis of Deviance Table

Model 1: cor_calcium ~ as.integer(age.q4)
Model 2: cor_calcium ~ age.q4
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       504     639.49
2       502     638.83  2  0.66029   0.7188
```

This suggests that there is no appreciable departure from linearity!

2. Assessing Assumptions

Extra Practice

Examine the grouped smooth approach for SBP.

❑ Regress coronary calcium on the 4 quartiles of SBP.

❑ Do the beta estimates for the slopes appear to be increasing linearly?

❑ Plot the change in logit corresponding to each of the quartiles, vs. the mean of the values in each quartile.

2. Assessing Assumptions 22

LOESS (Locally-Estimated) Smoothing

Strategy: Similar to grouped smooth, but instead of using discrete categories, use a moving window/band.

- Calculate the logit($\hat{\pi}$) for each point in the dataset, using a weighted average regression of adjacent points (weighted by distance from the current point).

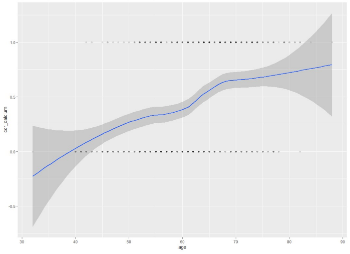
22

2. Assessing Assumptions 23

This is a graph of the relationship between age and predicted probability of coronary calcium, using the LOESS smoother.

This assesses the relationship between **age and the predicted probability of cor_calcium**.

Therefore…



23

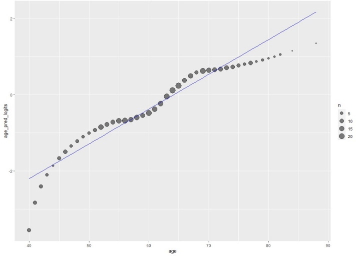
2. Assessing Assumptions 24

Instead, we want to examine the relationship between age and the <u>logit</u> of the probability of coronary calcium.

Note that this tells you the predicted logit across X values.

The LOESS smoother can be sensitive to the actual data. Therefore, it may pick up small departures from linearity.

This relationship between age and the logit of coronary calcium appears relatively linear.



24

2. Assessing Assumptions 25

Extra Practice

Examine the LOESS approach for SBP.

❑ Find the predicted probabilities and logits of coronary calcium over the values of SBP.

❑ Plot the predicted logit over the values of SBP.

25

2. Assessing Assumptions 26

Fractional Polynomials

Strategy: Find a transformation of X (e.g., log(X), X^2) that fits the data best.

- We have learned about fractional polynomials in Week 7, and the approach can be used here to examine the linearity assumption.

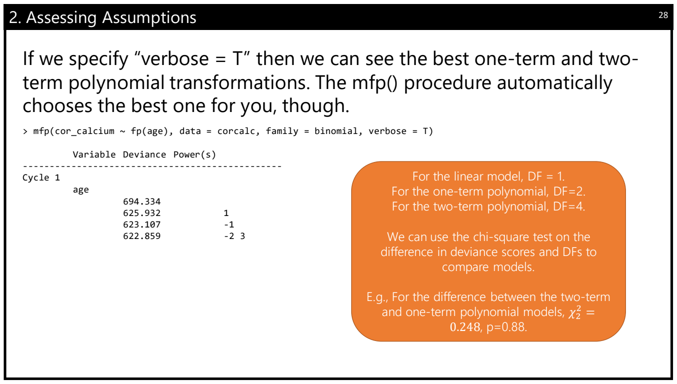
26

2. Assessing Assumptions 27

Here we see that there is no transformation to age would improve the model fit.

```
> mfp(cor_calcium ~ fp(age), data = corcalc, family = binomial)


Deviance table:
                   Resid. Dev
Null model         694.3336
Linear model       625.9316
Final model        625.9316

Fractional polynomials:
    df.initial select alpha df.final power1 power2
age          4      1  0.05        1      1      .

Transformations of covariates:
           formula
age I((age/100)^1)

Rescaled coefficients:
Intercept      age.1
 -5.26382    0.08203

Degrees of Freedom: 505 Total (i.e. Null);  504 Residual
Null Deviance:        694.3
Residual Deviance: 625.9    AIC: 629.9
```

27



2. Assessing Assumptions

If we specify “verbose = T” then we can see the best one-term and two-term polynomial transformations. The mfp() procedure automatically chooses the best one for you, though.

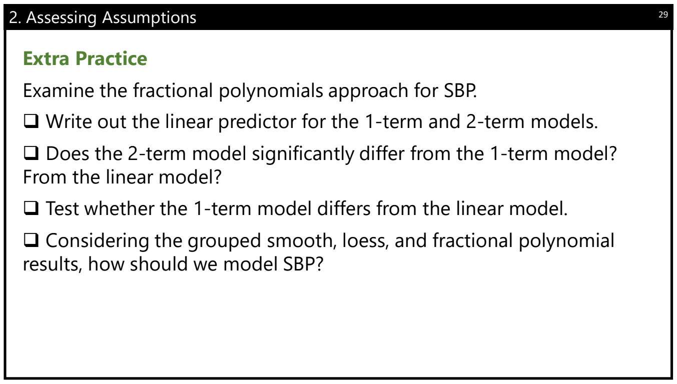
```
> mfp(cor_calcium ~ fp(age), data = corcalc, family = binomial, verbose = T)

        Variable Deviance  Power(s)
------------------------------------------------
Cycle 1
        age
                694.334
                625.932         1
                623.107         -1
                622.859         -2 3
```

For the linear model, DF = 1.
For the one-term polynomial, DF=2.
For the two-term polynomial, DF=4.

We can use the chi-square test on the difference in deviance scores and DFs to compare models.

E.g., For the difference between the two-term and one-term polynomial models, $\chi^2_2 = 0.248$, p=0.88.

28



2. Assessing Assumptions

Extra Practice

Examine the fractional polynomials approach for SBP.

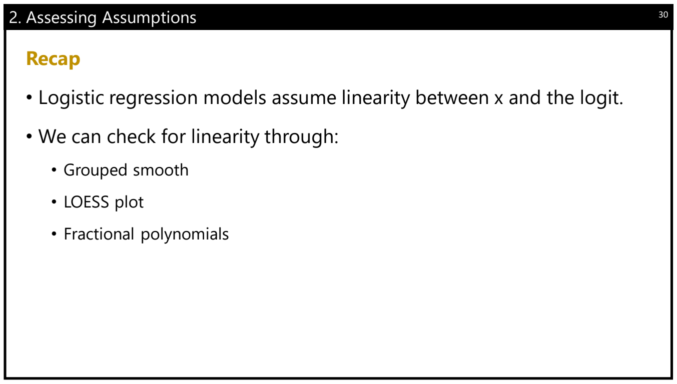
❑ Write out the linear predictor for the 1-term and 2-term models.

❑ Does the 2-term model significantly differ from the 1-term model? From the linear model?

❑ Test whether the 1-term model differs from the linear model.

❑ Considering the grouped smooth, loess, and fractional polynomial results, how should we model SBP?

29

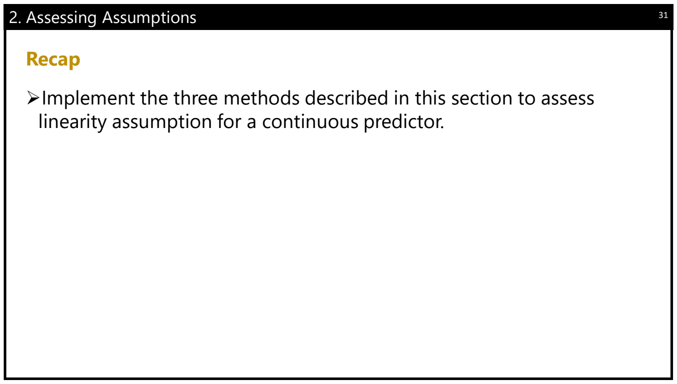


2. Assessing Assumptions

Recap

- Logistic regression models assume linearity between x and the logit.
- We can check for linearity through:
  - Grouped smooth
  - LOESS plot
  - Fractional polynomials

30

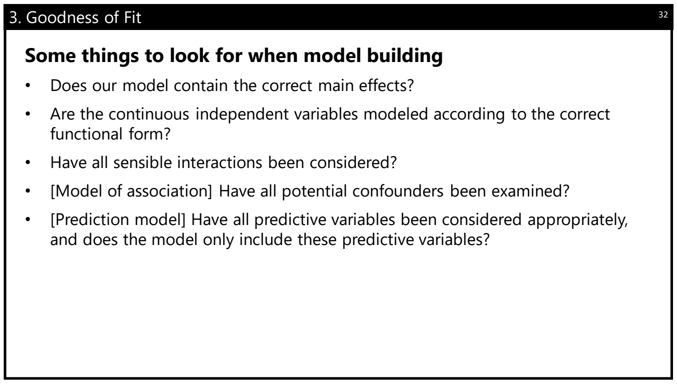


2. Assessing Assumptions 31

Recap

➢Implement the three methods described in this section to assess linearity assumption for a continuous predictor.

31

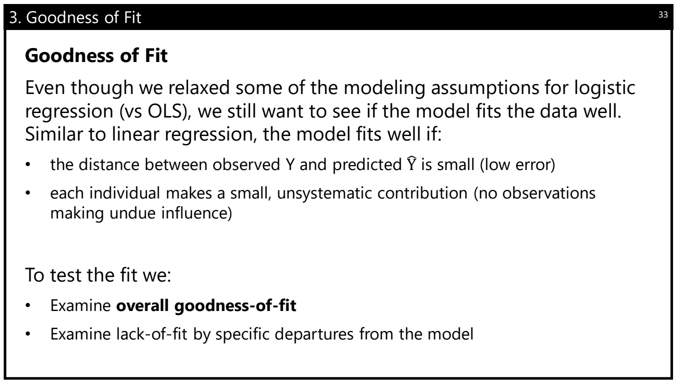


3. Goodness of Fit 32

Some things to look for when model building

- Does our model contain the correct main effects?
- Are the continuous independent variables modeled according to the correct functional form?
- Have all sensible interactions been considered?
- [Model of association] Have all potential confounders been examined?
- [Prediction model] Have all predictive variables been considered appropriately, and does the model only include these predictive variables?

32



3. Goodness of Fit 33

Goodness of Fit

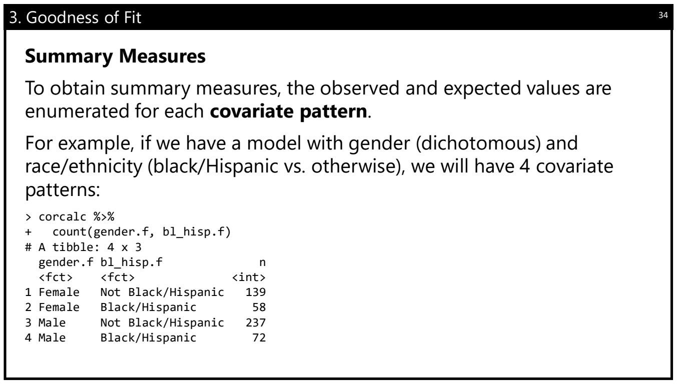
Even though we relaxed some of the modeling assumptions for logistic regression (vs OLS), we still want to see if the model fits the data well. Similar to linear regression, the model fits well if:

- the distance between observed Y and predicted $\hat{Y}$ is small (low error)
- each individual makes a small, unsystematic contribution (no observations making undue influence)

To test the fit we:

- Examine **overall goodness-of-fit**
- Examine lack-of-fit by specific departures from the model

33



3. Goodness of Fit

34

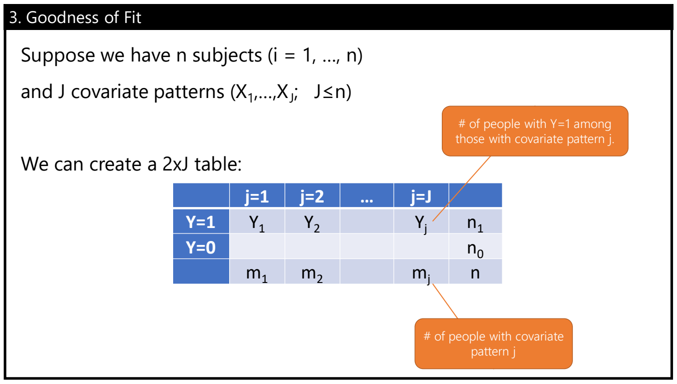
Summary Measures

To obtain summary measures, the observed and expected values are enumerated for each **covariate pattern**.

For example, if we have a model with gender (dichotomous) and race/ethnicity (black/Hispanic vs. otherwise), we will have 4 covariate patterns:

```
> corcalc %>%
+   count(gender.f, bl_hisp.f)
# A tibble: 4 x 3
  gender.f bl_hisp.f              n
  <fct>    <fct>              <int>
1 Female   Not Black/Hispanic   139
2 Female   Black/Hispanic        58
3 Male     Not Black/Hispanic   237
4 Male     Black/Hispanic        72
```

34



3. Goodness of Fit

Suppose we have n subjects (i = 1, …, n)

and J covariate patterns ($X_1,…,X_J$;   J≤n)

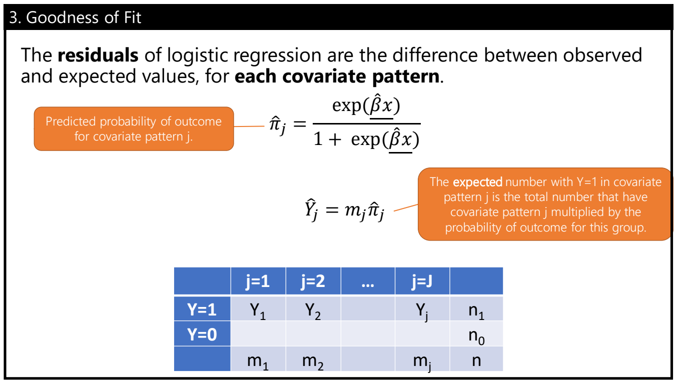
We can create a 2xJ table:

	j=1	j=2	…	j=J	
Y=1	Y_1	Y_2		Y_j	n_1
Y=0					n_0
	m_1	m_2		m_j	n

# of people with Y=1 among those with covariate pattern j.

# of people with covariate pattern j

35



3. Goodness of Fit

The **residuals** of logistic regression are the difference between observed and expected values, for **each covariate pattern**.

Predicted probability of outcome for covariate pattern j.

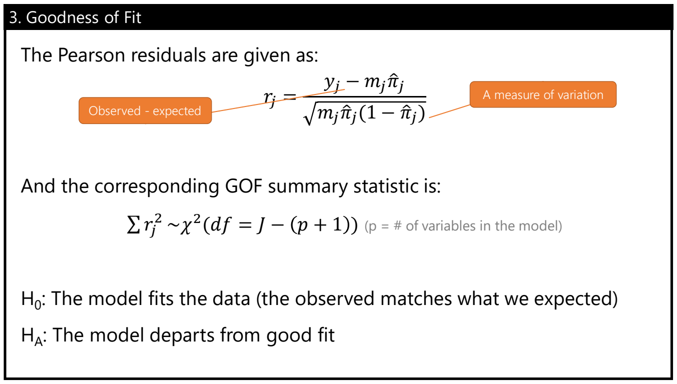
$$\hat{\pi}_j = \frac{\exp(\underline{\hat{\beta} x})}{1 + \exp(\underline{\hat{\beta} x})}$$

$$\hat{Y}_j = m_j \hat{\pi}_j$$

The **expected** number with Y=1 in covariate pattern j is the total number that have covariate pattern j multiplied by the probability of outcome for this group.

	j=1	j=2	…	j=J	
Y=1	Y_1	Y_2		Y_j	n_1
Y=0					n_0
	m_1	m_2		m_j	n

36



3. Goodness of Fit

The Pearson residuals are given as:

$$r_j = \frac{y_j - m_j\hat{\pi}_j}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}}$$

Observed - expected

A measure of variation

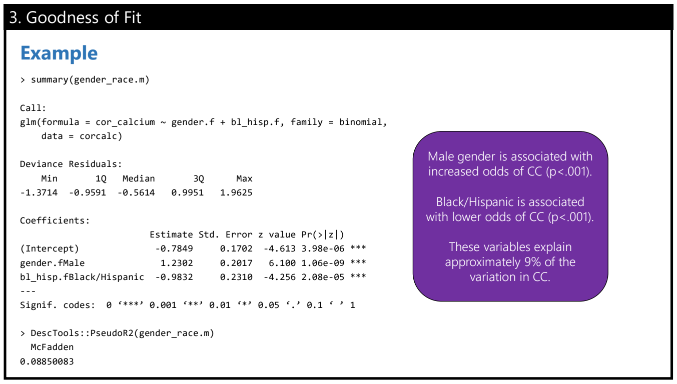
And the corresponding GOF summary statistic is:

$$\sum r_j^2 \sim \chi^2(df = J-(p+1))$$ (p = # of variables in the model)

H_0: The model fits the data (the observed matches what we expected)

H_A: The model departs from good fit

37



3. Goodness of Fit

Example

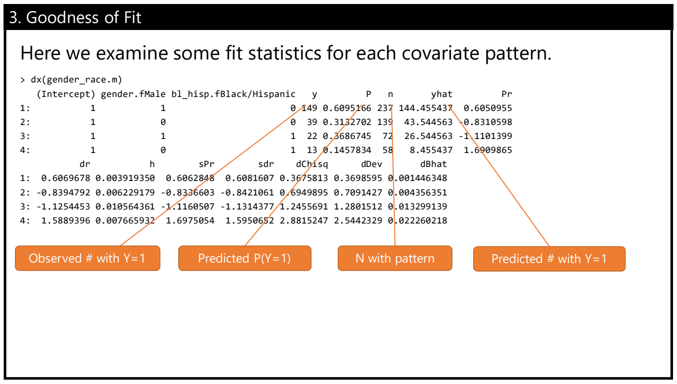
```
> summary(gender_race.m)

Call:
glm(formula = cor_calcium ~ gender.f + bl_hisp.f, family = binomial,
    data = corcalc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3714  -0.9591  -0.5614   0.9951   1.9625

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)              -0.7849     0.1702  -4.613 3.98e-06 ***
gender.fMale              1.2302     0.2017   6.100 1.06e-09 ***
bl_hisp.fBlack/Hispanic  -0.9832     0.2310  -4.256 2.08e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> DescTools::PseudoR2(gender_race.m)
  McFadden
0.08850083
```

Male gender is associated with increased odds of CC (p<.001).

Black/Hispanic is associated with lower odds of CC (p<.001).

These variables explain approximately 9% of the variation in CC.

38



3. Goodness of Fit

Here we examine some fit statistics for each covariate pattern.

```
> dx(gender_race.m)
   (Intercept) gender.fMale bl_hisp.fBlack/Hispanic   y         P   n       yhat         Pr
1:           1            1                       0 149 0.6095166 237 144.455437  0.6050955
2:           1            0                       0  39 0.3132702 139  43.544563 -0.8310598
3:           1            1                       1  22 0.3686745  72  26.544563 -1.1101399
4:           1            0                       1  13 0.1457834  58   8.455437  1.6909865
          dr           h        sPr        sdr    dChisq      dDev       dBhat
1:  0.6069678 0.003919350  0.6062848  0.6081607 0.3675813 0.3698595 0.001446348
2: -0.8394792 0.006229179 -0.8336603 -0.8421061 0.6949895 0.7091427 0.004356351
3: -1.1254453 0.010564361 -1.1160507 -1.1314377 1.2455691 1.2801512 0.013299139
4:  1.5889396 0.007665932  1.6975054  1.5950652 2.8815247 2.5442329 0.022260218
```

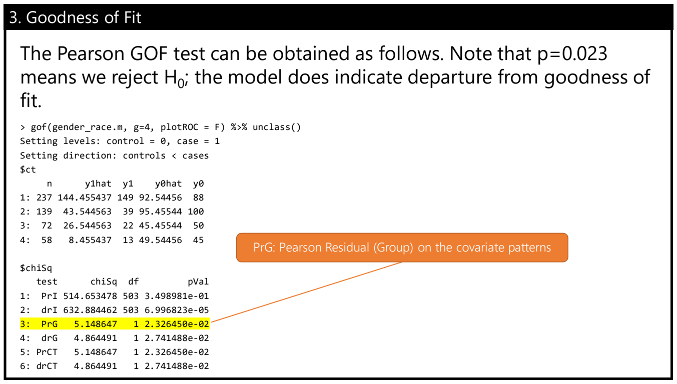
Observed # with Y=1

Predicted P(Y=1)

N with pattern

Predicted # with Y=1

39



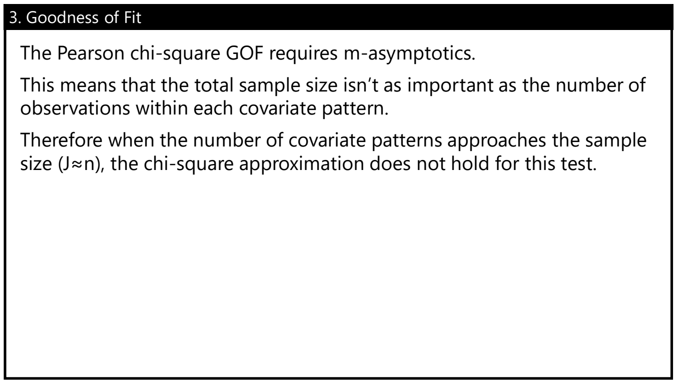
3. Goodness of Fit

The Pearson GOF test can be obtained as follows. Note that p=0.023 means we reject H_0; the model does indicate departure from goodness of fit.

```
> gof(gender_race.m, g=4, plotROC = F) %>% unclass()
Setting levels: control = 0, case = 1
Setting direction: controls < cases
$ct
     n      y1hat  y1    y0hat  y0
1: 237 144.455437 149 92.54456  88
2: 139  43.544563  39 95.45544 100
3:  72  26.544563  22 45.45544  50
4:  58   8.455437  13 49.54456  45

$chiSq
   test      chiSq  df         pVal
1:  PrI 514.653478 503 3.498981e-01
2:  drI 632.884462 503 6.996823e-05
3:  PrG   5.148647   1 2.326450e-02
4:  drG   4.864491   1 2.741488e-02
5: PrCT   5.148647   1 2.326450e-02
6: drCT   4.864491   1 2.741488e-02
```

PrG: Pearson Residual (Group) on the covariate patterns

40



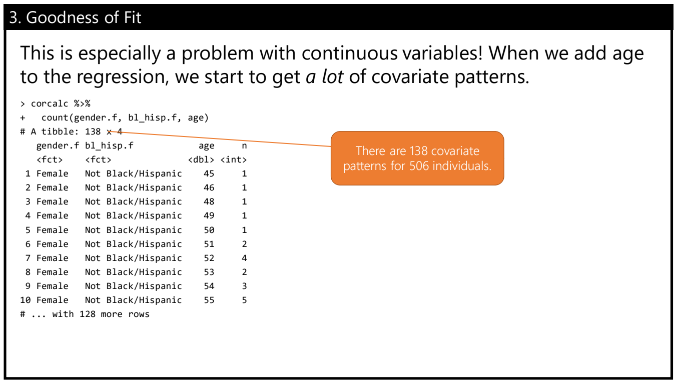
3. Goodness of Fit

The Pearson chi-square GOF requires m-asymptotics.

This means that the total sample size isn't as important as the number of observations within each covariate pattern.

Therefore when the number of covariate patterns approaches the sample size (J≈n), the chi-square approximation does not hold for this test.

41



3. Goodness of Fit

This is especially a problem with continuous variables! When we add age to the regression, we start to get *a lot* of covariate patterns.

```
> corcalc %>%
+   count(gender.f, bl_hisp.f, age)
# A tibble: 138 x 4
   gender.f bl_hisp.f             age     n
   <fct>    <fct>               <dbl> <int>
 1 Female   Not Black/Hispanic     45     1
 2 Female   Not Black/Hispanic     46     1
 3 Female   Not Black/Hispanic     48     1
 4 Female   Not Black/Hispanic     49     1
 5 Female   Not Black/Hispanic     50     1
 6 Female   Not Black/Hispanic     51     2
 7 Female   Not Black/Hispanic     52     4
 8 Female   Not Black/Hispanic     53     2
 9 Female   Not Black/Hispanic     54     3
10 Female   Not Black/Hispanic     55     5
# ... with 128 more rows
```

There are 138 covariate patterns for 506 individuals.

42

3. Goodness of Fit

Hosmer-Lemeshow GOF Test

An alternative to the Pearson GOF test that “fixes” the problem of having too many covariate patterns.

How?

1. Collapse the J covariate patterns into g groups (g<J, and fix g<<n). Then calculate the observed and expected frequencies.
2. Obtain the predicted probabilities, $\hat{\pi}_j$, for each covariate pattern j.
3. Order the j columns (covariate patterns) from lowest to highest predicted probabilities.
4. Collapse the J columns into deciles of risk (g=10)
5. Calculate expected values for each of the 10 categories (sum over all subjects in the cells with Y=1 or in cells with Y=0).
6. Perform chi-square test and compare to a χ^2 with g-2 degrees of freedom.

43

3. Goodness of Fit

```
> hoslem.test(gender_race_age.m$y, fitted(gender_race_age.m), g=10)

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  gender_race_age.m$y, fitted(gender_race_age.m)
X-squared = 6.3405, df = 8, p-value = 0.6092

> hoslem.test(gender_race_age.m$y, fitted(gender_race_age.m), g=10) %>%
+   {cbind(.$observed, .$expected)}
              y0 y1     yhat0     yhat1
[0.031,0.115] 47  4 46.695674  4.304326
(0.115,0.181] 43  8 43.233451  7.766549
(0.181,0.25]  44 11 42.914692 12.085308
(0.25,0.312]  38 13 36.245566 14.754434
(0.312,0.416] 29 16 28.553477 16.446523
(0.416,0.496] 25 29 29.565518 24.434482
(0.496,0.609] 25 24 21.706068 27.293932
(0.609,0.713] 14 39 17.400573 35.599427
(0.713,0.821]  9 40 10.884210 38.115790
(0.821,0.967]  9 39  5.800773 42.199227
```

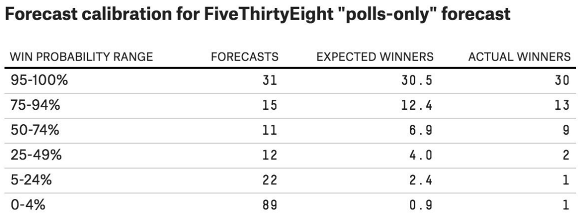
There is no evidence of lack of fit (p=0.61).

The range of predicted probabilities is split into 10 quantiles. Within each quantile, we calculate how many observations with Y=0 and with Y=1 we expect, and compare that to how many we would observe. You can see that the observed closely matches the expected.

44

3. Goodness of Fit

This kind of test can be used to make sure that prediction models are calibrated correctly.

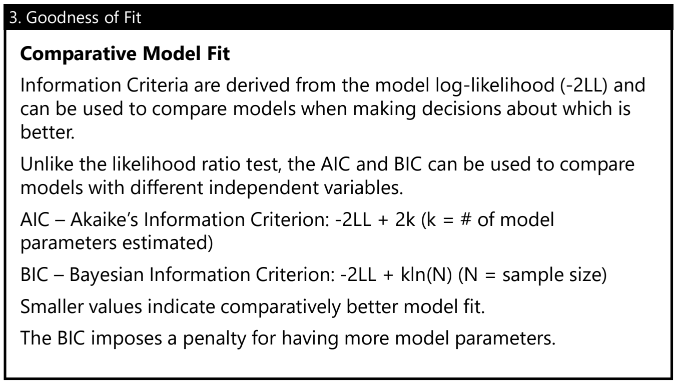


Forecast calibration for FiveThirtyEight "polls-only" forecast

WIN PROBABILITY RANGE	FORECASTS	EXPECTED WINNERS	ACTUAL WINNERS
95-100%	31	30.5	30
75-94%	15	12.4	13
50-74%	11	6.9	9
25-49%	12	4.0	2
5-24%	22	2.4	1
0-4%	89	0.9	1

Cole Fitzpatrick @colefitzpatrick · May 11, 2016
Replying to @NateSilver538
@NateSilver538 Ah, the Hosmer-Lemeshow test.

45



3. Goodness of Fit

Comparative Model Fit

Information Criteria are derived from the model log-likelihood (-2LL) and can be used to compare models when making decisions about which is better.

Unlike the likelihood ratio test, the AIC and BIC can be used to compare models with different independent variables.

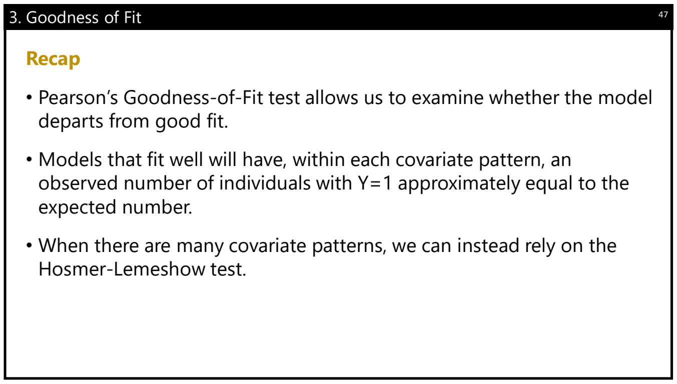
AIC – Akaike's Information Criterion: -2LL + 2k (k = # of model parameters estimated)

BIC – Bayesian Information Criterion: -2LL + kln(N) (N = sample size)

Smaller values indicate comparatively better model fit.

The BIC imposes a penalty for having more model parameters.

46

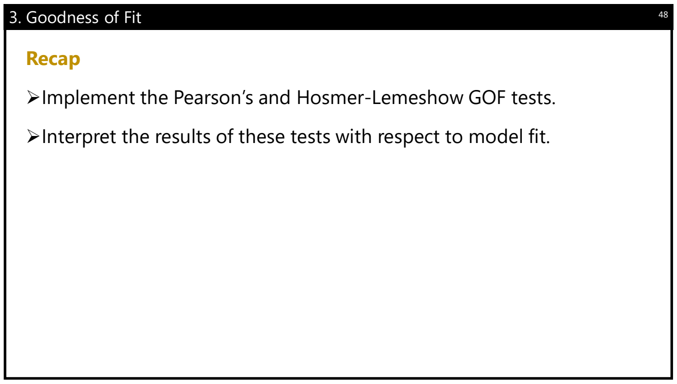


3. Goodness of Fit 47

Recap

- Pearson's Goodness-of-Fit test allows us to examine whether the model departs from good fit.
- Models that fit well will have, within each covariate pattern, an observed number of individuals with Y=1 approximately equal to the expected number.
- When there are many covariate patterns, we can instead rely on the Hosmer-Lemeshow test.

47



3. Goodness of Fit 48

Recap

➢Implement the Pearson's and Hosmer-Lemeshow GOF tests.

➢Interpret the results of these tests with respect to model fit.

48

4. Diagnostics 49

Diagnostics

As with linear regression, we need to check:

- Collinearity
- Leverage
- Influence

49

4. Diagnostics 50

Collinearity

We can check for collinearity as we normally would with OLS regression.

```
> DescTools::VIF(gender_race_age.m)
 gender.f bl_hisp.f       age
 1.160517  1.003035  1.163754
```

There is no evidence of collinearity. The largest VIF is 1.16, far below 10.

50

4. Diagnostics 51

Leverage

Recall, leverage indicates observations that have the potential to be influential because they are far from the average value of a covariate.

In linear regression, leverage values are obtained from the hat matrix: H = X(X'X)-1X'.

In logistic regression, H = V1/2 X(X'VX)-1X'V1/2 , where V is a JxJ diagonal matrix with element $v_j = m_j \, \hat{\pi}(x_j)(1 - \hat{\pi}(x_j))$.

H is the leverage; the distance of covariate pattern X_j from the mean.

51

4. Diagnostics 52

Influence

An observation is influential when it has a high **residual** and a large value of **leverage**.

Influence is assessed by estimating the effect of deleting all subjects with a particular covariate pattern J.

We can see how this affects:

- The estimated coefficients (betas)
- The summary GOF measures

52

4. Diagnostics 53

Influence

We typically want to see the following plots:

- $\Delta\chi_j^2$ vs $\hat{\pi}_j$ (Change in Pearson GOF)
- ΔD_j vs $\hat{\pi}_j$ *(*Change in Deviance GOF)
- $\Delta\hat{\beta}_j$ vs $\hat{\pi}_j$ (Change in Cook's Distance)

53

4. Diagnostics 54

These values can be produced either:

- For each covariate pattern
- For each individual

```
dx(gender_race_age.m)
dx(gender_race_age.m, bycov = F)
```

54

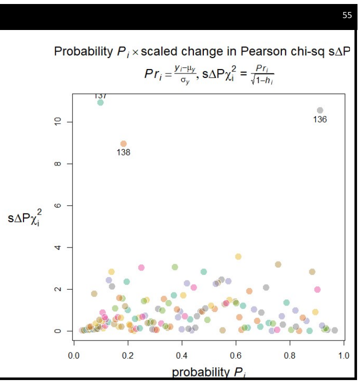
4. Diagnostics

55

$\Delta\chi_j^2$ vs $\hat{\pi}_j$

Poorly fit points will lie in the upper corners.

Assuming m-asymptotics, 4 is a crude approximation of the upper 95th percentile of the distribution of $\Delta\chi_j^2$.



55

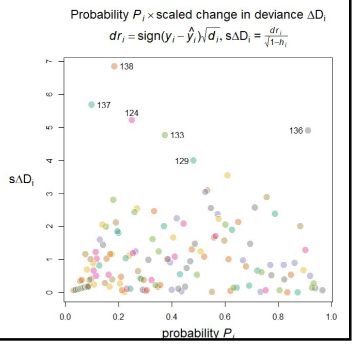
4. Diagnostics

56

ΔD_j vs $\hat{\pi}_j$

Poorly fit points will lie in the upper corners.

Assuming m-asymptotics, 4 is a crude approximation of the upper 95th percentile of the distribution of ΔD_j.



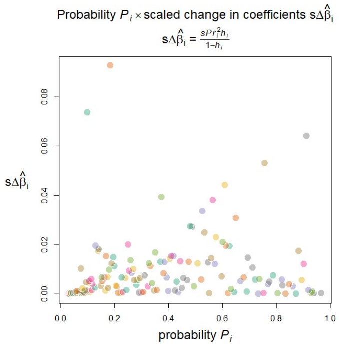
56

4. Diagnostics

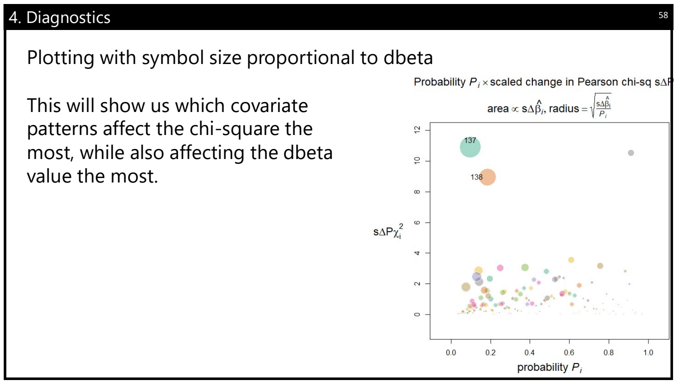
57

$\Delta\beta_j$ vs $\hat{\pi}_j$

Values above 1.0 indicate removal of the covariate pattern is associated with considerable changes to the parameter estimates.



57



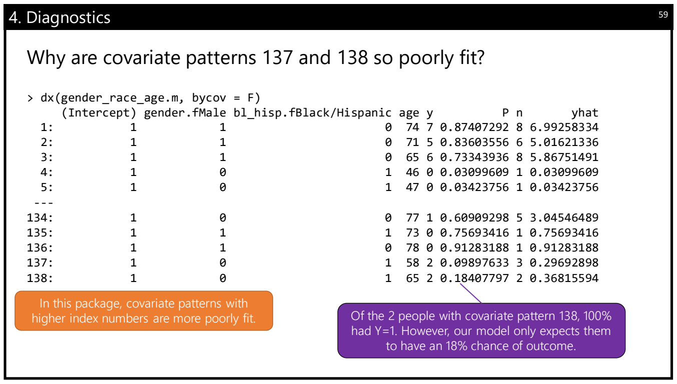
4. Diagnostics

58

Plotting with symbol size proportional to dbeta

This will show us which covariate patterns affect the chi-square the most, while also affecting the dbeta value the most.

58



4. Diagnostics

59

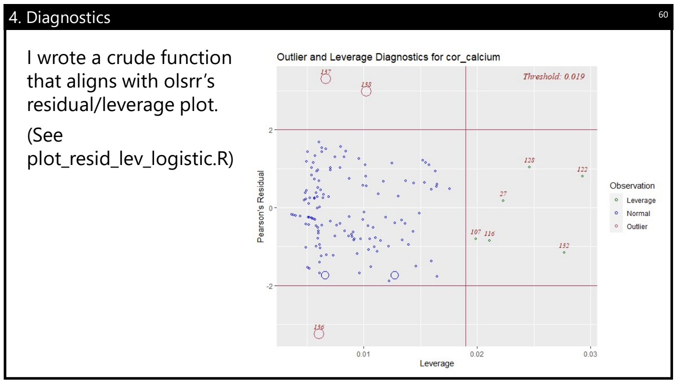
Why are covariate patterns 137 and 138 so poorly fit?

```
> dx(gender_race_age.m, bycov = F)
     (Intercept) gender.fMale bl_hisp.fBlack/Hispanic age y          P n       yhat
  1:           1            1                       0  74 7 0.87407292 8 6.99258334
  2:           1            1                       0  71 5 0.83603556 6 5.01621336
  3:           1            1                       0  65 6 0.73343936 8 5.86751491
  4:           1            0                       1  46 0 0.03099609 1 0.03099609
  5:           1            0                       1  47 0 0.03423756 1 0.03423756
 ---
134:           1            0                       0  77 1 0.60909298 5 3.04546489
135:           1            1                       1  73 0 0.75693416 1 0.75693416
136:           1            1                       0  78 0 0.91283188 1 0.91283188
137:           1            0                       1  58 2 0.09897633 3 0.29692898
138:           1            0                       1  65 2 0.18407797 2 0.36815594
```

In this package, covariate patterns with higher index numbers are more poorly fit.

Of the 2 people with covariate pattern 138, 100% had Y=1. However, our model only expects them to have an 18% chance of outcome.

59



4. Diagnostics

60

I wrote a crude function that aligns with olsrr's residual/leverage plot.

(See plot_resid_lev_logistic.R)

60

4. Diagnostics 61

What happens when we find problematic observations?

List the covariate pattern to see why the observation is influential.

You can delete these patterns and refit the model to determine the true effect of these observations on your $\hat{\beta}$ of interest.

Then decide:

- What is the reason for the outliers? If you delete them, you must have a valid reason to do so.
- Are the outlying patterns reasonable? Or are they due to a mistake?
- Is there a variable or set of variables you didn't include that would fix the model?

61

4. Diagnostics 62

What if there are multiple suspect patterns?

Check the following:

- Did you use the correct link?
- Did you omit an important predictor or interaction?
- Are the covariates on the proper scale?
- Is there "extra-binomial variation"? (more or less variation in predicted probabilities than expected under the binomial model; can occur when observations are clustered)

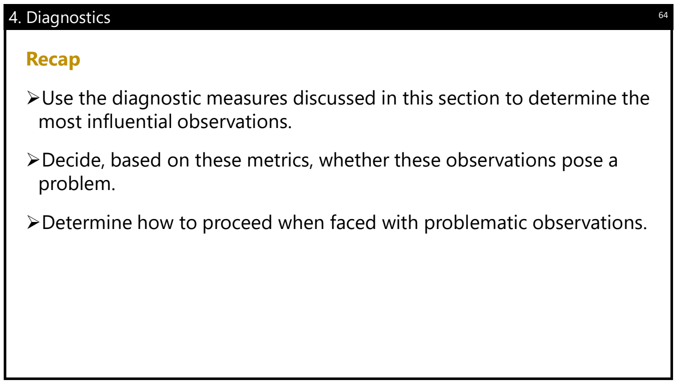
62

4. Diagnostics 63

Recap

- An examination of the change in Pearson's GOF, Deviance GOF, and betas can help identify covariate patterns that are poorly fit.

63

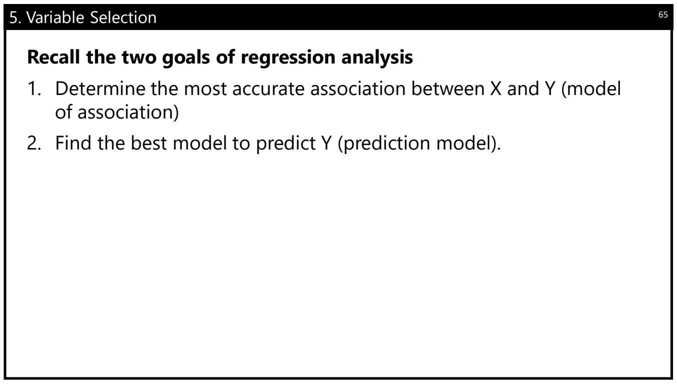


4. Diagnostics 64

Recap

- ➢Use the diagnostic measures discussed in this section to determine the most influential observations.
- ➢Decide, based on these metrics, whether these observations pose a problem.
- ➢Determine how to proceed when faced with problematic observations.

64

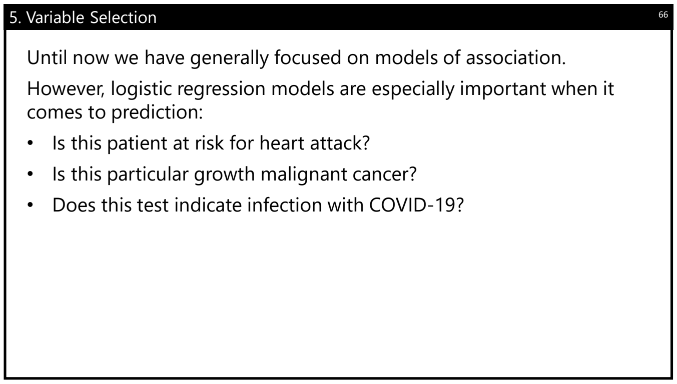


5. Variable Selection 65

Recall the two goals of regression analysis

1. Determine the most accurate association between X and Y (model of association)
2. Find the best model to predict Y (prediction model).

65



5. Variable Selection 66

Until now we have generally focused on models of association.

However, logistic regression models are especially important when it comes to prediction:

- Is this patient at risk for heart attack?
- Is this particular growth malignant cancer?
- Does this test indicate infection with COVID-19?

66

5. Variable Selection

67

Example

Can we use characteristics of the mother in order to predict low birth weight?

```
> lbw %>%
+   select(LOW, AGE, LWT, RACE, SMOKE, PTL, HT, UI, FTV) %>%
+   psych::describe()
       vars   n   mean    sd median trimmed   mad min max range skew kurtosis   se
LOW*      1 189   1.31  0.46      1    1.27  0.00   1   2     1 0.80    -1.36 0.03
AGE       2 189  23.24  5.30     23   22.90  5.93  14  45    31 0.71     0.53 0.39
LWT       3 189 129.81 30.58    121  126.07 20.76  80 250   170 1.38     2.25 2.22
RACE*     4 189   1.85  0.92      1    1.81  0.00   1   3     2 0.31    -1.75 0.07
SMOKE*    5 189   1.39  0.49      1    1.37  0.00   1   2     1 0.44    -1.82 0.04
PTL       6 189   0.20  0.49      0    0.08  0.00   0   3     3 2.76     8.17 0.04
HT*       7 189   1.06  0.24      1    1.00  0.00   1   2     1 3.55    10.67 0.02
UI*       8 189   1.15  0.36      1    1.07  0.00   1   2     1 1.97     1.87 0.03
FTV       9 189   0.79  1.06      0    0.62  0.00   0   6     6 1.56     3.00 0.08
```

67

5. Variable Selection

68

When faced with several possible predictive variables, it can be cumbersome to manually arrive at a good model.

Automatic selection procedures (while criticized for being too “hands-off”) provide a way to assess which variables may be important.

Selection Algorithms

- Best Subsets
- Backward Elimination
- Forward Selection
- Stepwise Selection

68

5. Variable Selection

69

Traditionally, these selection algorithms were based on p-values.

i.e., add the most significant variables to the model according to their p-value until they’re no longer significant.

Recently, there has been a push to stop using p-values as a criterion for model inclusion/exclusion and instead turn to other measures, such as R^2 or the information criteria (AIC/BIC/etc.).

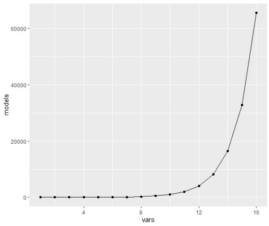
69

5. Variable Selection 70

Best Subsets

For K variates under consideration, assess the fit of all models with *k=1, 2, 3, …, K* variables included in the model.

- The best subset is chosen using <u>some</u> criteria (Information Criterion, R^2, Mallow's C_p, etc.)
- This approach is computationally intensive, as it requires fitting 2^K-1 models.



70

5. Variable Selection 71

Finding the best subset of predictors

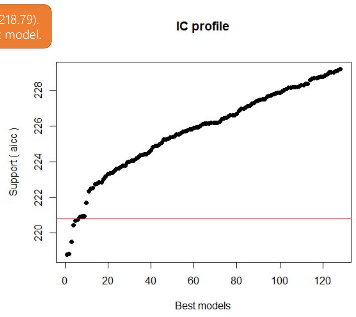
```
best_subset_low <-
glmulti(LOW ~ AGE + LWT + RACE + SMOKE + PTL + HT + UI + FTV, data=lbw,
        level=1, family = binomial, crit="aicc", confsetsize=128)

> print(best_subset_low)
glmulti.analysis
Method: h / Fitting: glm / IC used: aicc
Level: 1 / Marginality: FALSE
From 128 models:
Best IC: 218.785587197454
Best model:
[1] "LOW ~ 1 + RACE + SMOKE + HT + UI + LWT + PTL"
Evidence weight: 0.103340083255988
Worst IC: 229.190066247819
6 models within 2 IC units.
80 models to reach 95% of evidence weight.
```

Level=1 considers main effects. Interactions can be considered by specifying level=2.

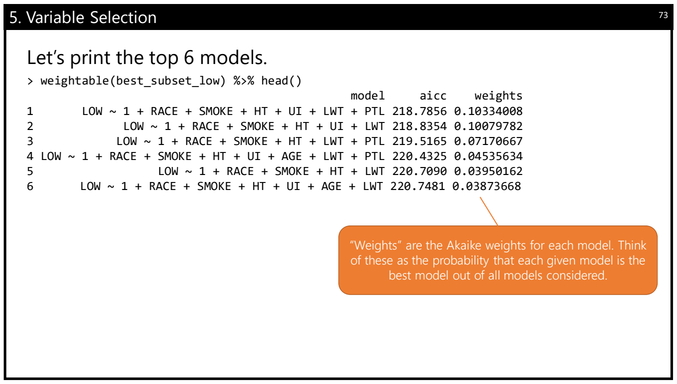
71

5. Variable Selection 72

The best model has the lowest AICC (218.79). 6 models are within 2 units of the best model.



72



5. Variable Selection

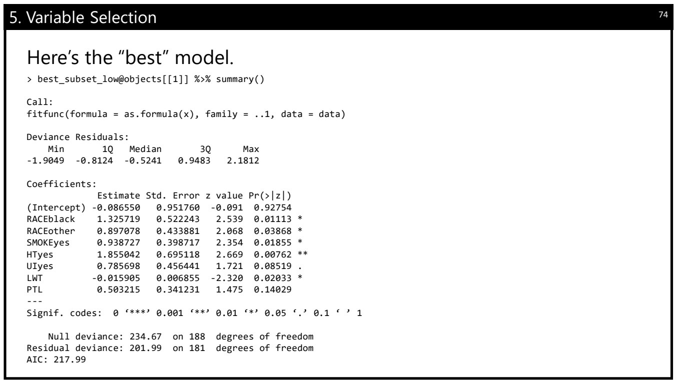
73

Let’s print the top 6 models.

```
> weightable(best_subset_low) %>% head()
                                                   model     aicc    weights
1       LOW ~ 1 + RACE + SMOKE + HT + UI + LWT + PTL 218.7856 0.10334008
2             LOW ~ 1 + RACE + SMOKE + HT + UI + LWT 218.8354 0.10079782
3            LOW ~ 1 + RACE + SMOKE + HT + LWT + PTL 219.5165 0.07170667
4 LOW ~ 1 + RACE + SMOKE + HT + UI + AGE + LWT + PTL 220.4325 0.04535634
5                  LOW ~ 1 + RACE + SMOKE + HT + LWT 220.7090 0.03950162
6       LOW ~ 1 + RACE + SMOKE + HT + UI + AGE + LWT 220.7481 0.03873668
```

“Weights” are the Akaike weights for each model. Think of these as the probability that each given model is the best model out of all models considered.

73



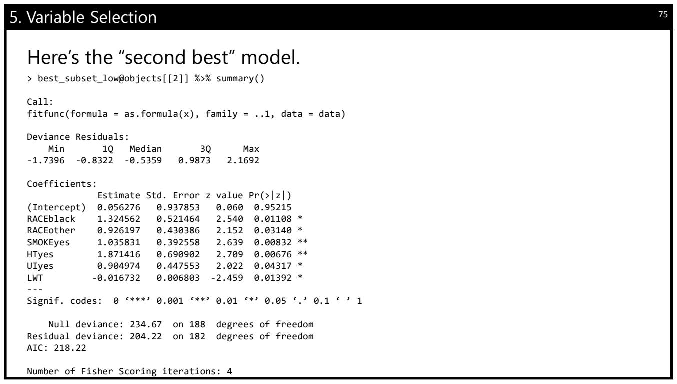
5. Variable Selection

74

Here’s the “best” model.

```
> best_subset_low@objects[[1]] %>% summary()

Call:
fitfunc(formula = as.formula(x), family = ..1, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9049  -0.8124  -0.5241   0.9483   2.1812

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.086550   0.951760  -0.091  0.92754
RACEblack    1.325719   0.522243   2.539  0.01113 *
RACEother    0.897078   0.433881   2.068  0.03868 *
SMOKEyes     0.938727   0.398717   2.354  0.01855 *
HTyes        1.855042   0.695118   2.669  0.00762 **
UIyes        0.785698   0.456441   1.721  0.08519 .
LWT         -0.015905   0.006855  -2.320  0.02033 *
PTL          0.503215   0.341231   1.475  0.14029
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 201.99  on 181  degrees of freedom
AIC: 217.99
```

74



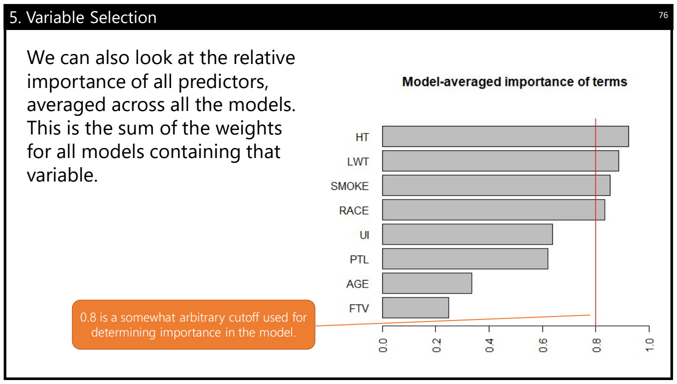
5. Variable Selection

75

Here’s the “second best” model.

```
> best_subset_low@objects[[2]] %>% summary()

Call:
fitfunc(formula = as.formula(x), family = ..1, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7396  -0.8322  -0.5359   0.9873   2.1692

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.056276   0.937853   0.060  0.95215
RACEblack    1.324562   0.521464   2.540  0.01108 *
RACEother    0.926197   0.430386   2.152  0.03140 *
SMOKEyes     1.035831   0.392558   2.639  0.00832 **
HTyes        1.871416   0.690902   2.709  0.00676 **
UIyes        0.904974   0.447553   2.022  0.04317 *
LWT         -0.016732   0.006803  -2.459  0.01392 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 204.22  on 182  degrees of freedom
AIC: 218.22

Number of Fisher Scoring iterations: 4
```

75

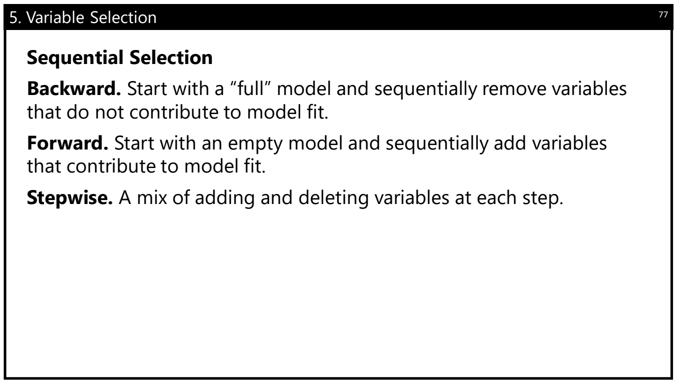


5. Variable Selection

We can also look at the relative importance of all predictors, averaged across all the models. This is the sum of the weights for all models containing that variable.

0.8 is a somewhat arbitrary cutoff used for determining importance in the model.

76



5. Variable Selection

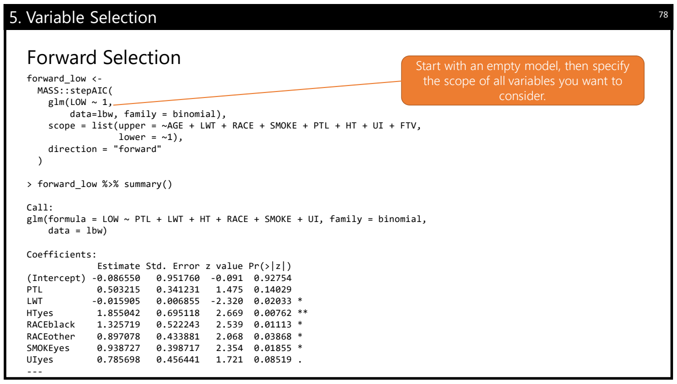
Sequential Selection

Backward. Start with a “full” model and sequentially remove variables that do not contribute to model fit.

Forward. Start with an empty model and sequentially add variables that contribute to model fit.

Stepwise. A mix of adding and deleting variables at each step.

77



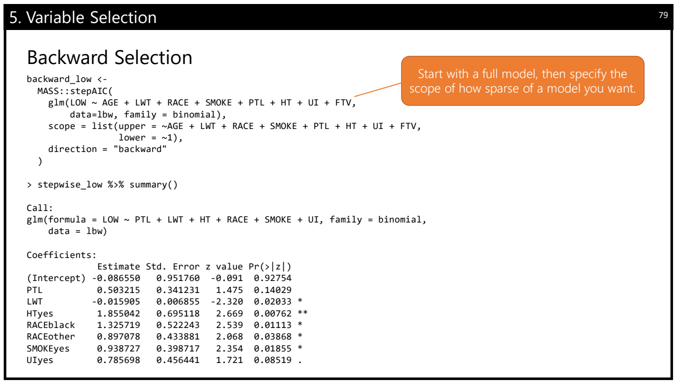
5. Variable Selection

Forward Selection

Start with an empty model, then specify the scope of all variables you want to consider.

```
forward_low <-
  MASS::stepAIC(
    glm(LOW ~ 1,
        data=lbw, family = binomial),
    scope = list(upper = ~AGE + LWT + RACE + SMOKE + PTL + HT + UI + FTV,
                 lower = ~1),
    direction = "forward"
  )

> forward_low %>% summary()

Call:
glm(formula = LOW ~ PTL + LWT + HT + RACE + SMOKE + UI, family = binomial,
    data = lbw)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.086550   0.951760  -0.091  0.92754
PTL          0.503215   0.341231   1.475  0.14029
LWT         -0.015905   0.006855  -2.320  0.02033 *
HTyes        1.855042   0.695118   2.669  0.00762 **
RACEblack    1.325719   0.522243   2.539  0.01113 *
RACEother    0.897078   0.433881   2.068  0.03868 *
SMOKEyes     0.938727   0.398717   2.354  0.01855 *
UIyes        0.785698   0.456441   1.721  0.08519 .
---
```

78



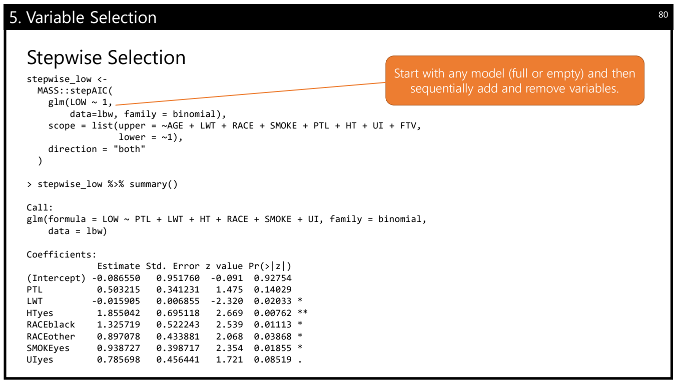
5. Variable Selection 79

Backward Selection

```
backward_low <-
  MASS::stepAIC(
    glm(LOW ~ AGE + LWT + RACE + SMOKE + PTL + HT + UI + FTV,
        data=lbw, family = binomial),
    scope = list(upper = ~AGE + LWT + RACE + SMOKE + PTL + HT + UI + FTV,
                 lower = ~1),
    direction = "backward"
  )

> stepwise_low %>% summary()

Call:
glm(formula = LOW ~ PTL + LWT + HT + RACE + SMOKE + UI, family = binomial,
    data = lbw)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.086550   0.951760  -0.091  0.92754
PTL          0.503215   0.341231   1.475  0.14029
LWT         -0.015905   0.006855  -2.320  0.02033 *
HTyes        1.855042   0.695118   2.669  0.00762 **
RACEblack    1.325719   0.522243   2.539  0.01113 *
RACEother    0.897078   0.433881   2.068  0.03868 *
SMOKEyes     0.938727   0.398717   2.354  0.01855 *
UIyes        0.785698   0.456441   1.721  0.08519 .
```

Start with a full model, then specify the scope of how sparse of a model you want.

79



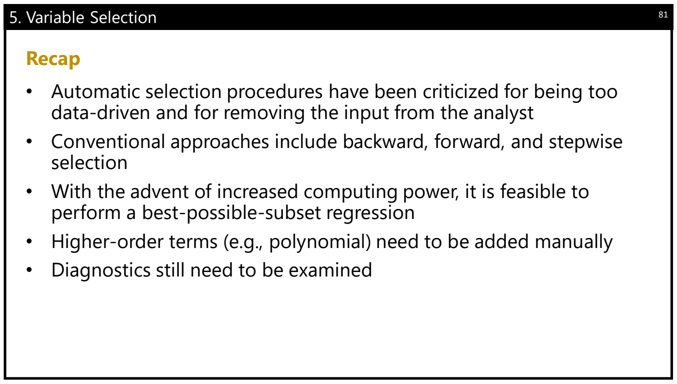
5. Variable Selection 80

Stepwise Selection

```
stepwise_low <-
  MASS::stepAIC(
    glm(LOW ~ 1,
        data=lbw, family = binomial),
    scope = list(upper = ~AGE + LWT + RACE + SMOKE + PTL + HT + UI + FTV,
                 lower = ~1),
    direction = "both"
  )

> stepwise_low %>% summary()

Call:
glm(formula = LOW ~ PTL + LWT + HT + RACE + SMOKE + UI, family = binomial,
    data = lbw)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.086550   0.951760  -0.091  0.92754
PTL          0.503215   0.341231   1.475  0.14029
LWT         -0.015905   0.006855  -2.320  0.02033 *
HTyes        1.855042   0.695118   2.669  0.00762 **
RACEblack    1.325719   0.522243   2.539  0.01113 *
RACEother    0.897078   0.433881   2.068  0.03868 *
SMOKEyes     0.938727   0.398717   2.354  0.01855 *
UIyes        0.785698   0.456441   1.721  0.08519 .
```

Start with any model (full or empty) and then sequentially add and remove variables.

80

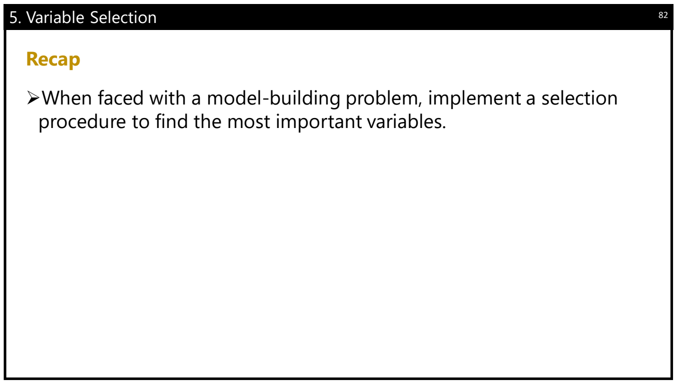


5. Variable Selection 81

Recap

- Automatic selection procedures have been criticized for being too data-driven and for removing the input from the analyst
- Conventional approaches include backward, forward, and stepwise selection
- With the advent of increased computing power, it is feasible to perform a best-possible-subset regression
- Higher-order terms (e.g., polynomial) need to be added manually
- Diagnostics still need to be examined

81

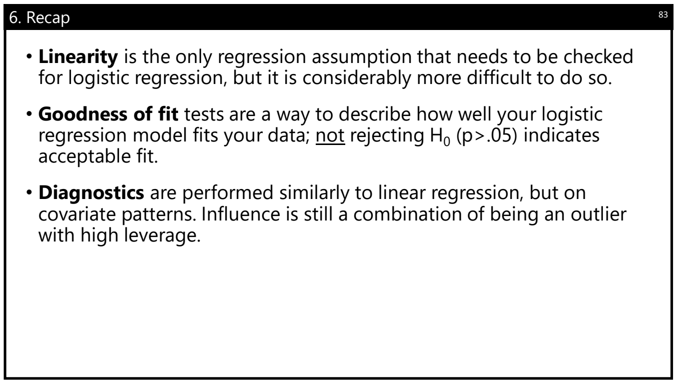


5. Variable Selection 82

Recap

➢When faced with a model-building problem, implement a selection procedure to find the most important variables.

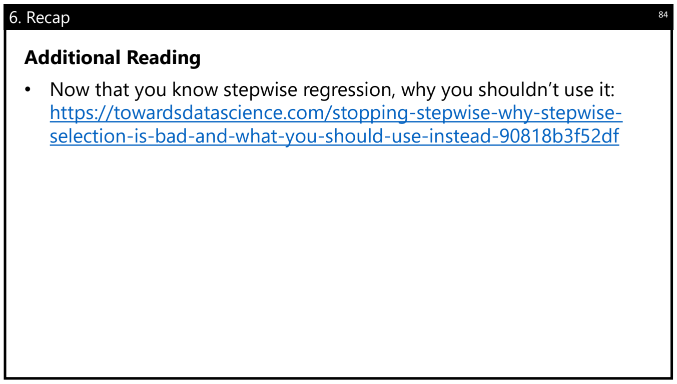
82



6. Recap 83

- **Linearity** is the only regression assumption that needs to be checked for logistic regression, but it is considerably more difficult to do so.
- **Goodness of fit** tests are a way to describe how well your logistic regression model fits your data; not rejecting H_0 ($p>.05$) indicates acceptable fit.
- **Diagnostics** are performed similarly to linear regression, but on covariate patterns. Influence is still a combination of being an outlier with high leverage.

83

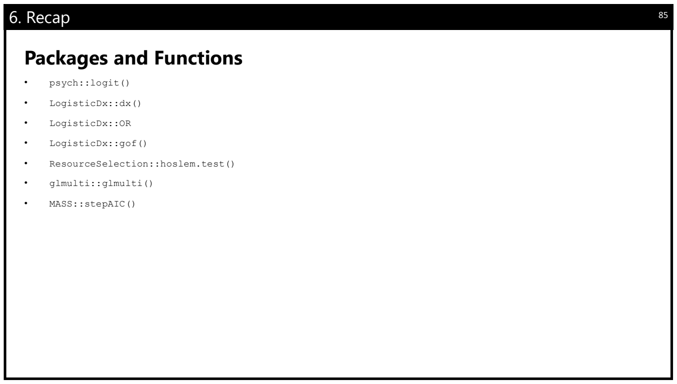


6. Recap 84

Additional Reading

- Now that you know stepwise regression, why you shouldn’t use it: https://towardsdatascience.com/stopping-stepwise-why-stepwise-selection-is-bad-and-what-you-should-use-instead-90818b3f52df

84



6. Recap

85

Packages and Functions

- `psych::logit()`
- `LogisticDx::dx()`
- `LogisticDx::OR`
- `LogisticDx::gof()`
- `ResourceSelection::hoslem.test()`
- `glmulti::glmulti()`
- `MASS::stepAIC()`

85