Outline

- Assessment of model fit
- Residuals
- Influence
- Model selection
- Prediction
Assessment of model fit – model deviance

The deviance of a fitted model compares the log-likelihood of the fitted model to the log-likelihood of a model with \( n \) parameters that fits the \( n \) observations perfectly. It can be shown that the likelihood of this saturated model is equal to 1 yielding a log-likelihood equal to 0. Therefore, the deviance for the logistic regression model is

\[
DEV = -2 \sum_{i=1}^{n} \left[ Y_i \log(\hat{\pi}_i) + (1 - Y_i) \log(1 - \hat{\pi}_i) \right],
\]

where \( \hat{\pi}_i \) is the fitted values for the \( i \)th observation. The smaller the deviance, the closer the fitted value is to the saturated model. The larger the deviance, the poorer the fit.
Sometimes, you will see a $\chi^2$ goodness of fit test based on the deviance, but this is inappropriate because the number of parameters in the saturated model is increasing at the same rate as $n$.

In the catheterization example,

$$\logit(\pi_i) = \beta_0 + \beta_1 sex_i$$ has deviance=3217,
$$\logit(\pi_i) = \beta_0 + \beta_1 age_i$$ has deviance=3153, and
$$\logit(\pi_i) = \beta_0 + \beta_1 cad.dur_i$$ has deviance=3131.

If we had to pick a model with only one predictor, which might we choose?
Hosmer-Lemeshow goodness of fit test

For this test,

\[ H_0 : E[Y] = \frac{\exp(X'\beta)}{1+\exp(X'\beta)} \]

\[ H_a : E[Y] \neq \frac{\exp(X'\beta)}{1+\exp(X'\beta)}. \]

To calculate the test statistic:

- Order the fitted values
- Group the fitted values into \( c \) classes (\( c \) is between 6 and 10) of roughly equal size
- Calculate the observed and expected number in each group
- Perform a \( \chi^2 \) goodness of fit test
Example with catheterization data:

\[ \logit(\pi_i) = \beta_0 + \beta_1 cad.dur_i + \beta_2 gender_i. \]

1. Order and group the fitted values

```r
> fi1 = fitted(glm11)
> fi1c = cut(fi1, br = c(0, quantile(fi1, p = seq(.1, .9, .1)), 1))
> table(fi1c)

   (0,0.371] (0.371,0.422] (0.422,0.426] (0.426,0.433] (0.433,0.442]
    239     323     180     227     198
   (0.442,0.47] (0.47,0.505] (0.505,0.555] (0.555,0.638] (0.638,1]
    236     230     233     237     229

> fi1c = cut(fi1, br = c(0, quantile(fi1, p = seq(.1, .9, .1)), 1), labels=F)
> table(fi1c)

   1   2   3   4   5   6   7   8   9  10
   239  323  180  227  198  236  230  233  237  229
```
2. Calculate the observed and expected values in each group

```r
E <- matrix(0, nrow=10, ncol=2)
O <- matrix(0, nrow=10, ncol=2)
for(j in 1:10) {
  E[j,2] <- sum(fi1[fi1c==j])
  E[j,1] <- sum((1-fi1)[fi1c==j])
  O[j,2] <- sum(acath$tvdlm[fi1c==j])
  O[j,1] <- sum((1-acath$tvdlm)[fi1c==j])
}
```

<table>
<thead>
<tr>
<th></th>
<th>O</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yi</td>
<td>1-Yi</td>
<td>1-pi</td>
</tr>
<tr>
<td>1</td>
<td>145</td>
<td>94</td>
</tr>
<tr>
<td>2</td>
<td>219</td>
<td>104</td>
</tr>
<tr>
<td>3</td>
<td>110</td>
<td>70</td>
</tr>
<tr>
<td>4</td>
<td>131</td>
<td>96</td>
</tr>
<tr>
<td>5</td>
<td>111</td>
<td>87</td>
</tr>
<tr>
<td>6</td>
<td>123</td>
<td>113</td>
</tr>
<tr>
<td>7</td>
<td>111</td>
<td>119</td>
</tr>
<tr>
<td>8</td>
<td>95</td>
<td>138</td>
</tr>
<tr>
<td>9</td>
<td>90</td>
<td>147</td>
</tr>
<tr>
<td>10</td>
<td>68</td>
<td>161</td>
</tr>
</tbody>
</table>

BIOST 515, Lecture 14
3. Calculate $\chi^2$ statistic

$$X^2 = \sum_{j=1}^{c} \sum_{k=0}^{1} \frac{(O_{jk} - E_{jk})^2}{E_{jk}} \sim \chi^2_{c-2}$$

$$= 21.56 > 15.5 = \chi^2_{8,.95};$$

therefore, we reject $H_0$.

```r
> sum((O-E)^2/E)
[1] 21.55852
> 1-pchisq(sum((O-E)^2/E),8)
[1] 0.005802828```
Residuals

Residuals can be useful for identifying potential outliers (observations not well fit by the model) or misspecified models. We will look at two types of residuals

- Deviance residuals
- Partial residuals
The deviance residual is useful for determining if individual points are not well fit by the model.

The deviance residual for the $i$th observation is the signed square root of the contribution of the $i$th case to the sum for the model deviance, $DEV$. For the $i$th observation, it is given by

$$dev_i = \pm \left\{ -2 \left[ Y_i \log(\hat{\pi}_i) + (1 - Y_i) \log(1 - \hat{\pi}_i) \right] \right\}^{1/2},$$

where the sign is positive when $Y_i \geq \hat{\pi}_i$ and negative otherwise.

You can get the deviance residuals using the function $\text{residuals()}$ in $R$. 
Catheterization example

\[ \text{logit}(\pi_i) = \beta_0 + \beta_1 \text{cad.dur}_i + \beta_2 \text{gender}_i \]
Partial residuals

The **partial residual** is useful for assessing how the predictors should be transformed. For the $i$th observation, the partial residual for the $j$th predictor is

$$r_{ij} = \hat{\beta}_j X_{ij} + \frac{Y_i - \hat{\pi}_i}{\hat{\pi}_i (1 - \hat{\pi}_i)}.$$

This approach assumes additivity of predictors.
Influential observations

As in linear regression, we can use \textit{DFFITS} and \textit{DFBETAS} to identify influential observations.
The potentially influential observations we’ve identified are:

<table>
<thead>
<tr>
<th>sex</th>
<th>age</th>
<th>cad.dur</th>
<th>choleste</th>
<th>sigdz</th>
<th>tvdlm</th>
</tr>
</thead>
<tbody>
<tr>
<td>314</td>
<td>1</td>
<td>63</td>
<td>364</td>
<td>350</td>
<td>1</td>
</tr>
<tr>
<td>1239</td>
<td>1</td>
<td>61</td>
<td>349</td>
<td>250</td>
<td>1</td>
</tr>
</tbody>
</table>

As it turns out, these are the two most extreme observations in duration for males without severe coronary artery disease.
Model selection

As in simple linear regression, we can use AIC for model comparison or in a stepwise model selection routine. The same cautions and pros and cons apply.

> stepAIC(glm(tvdlm ~ sex * age * cad.dur, family = binomial, data = acath))
Start:  AIC = 3069.54
  tvdlm ~ sex * age * cad.dur

Df  Deviance      AIC
- sex:age:cad.dur 1 3055.2 3069.2
<none>            3053.5 3069.5

Step:  AIC = 3069.21
  tvdlm ~ sex + age + cad.dur + sex:age + sex:cad.dur + age:cad.dur

Df  Deviance      AIC
- sex:age      1 3055.3 3067.3
- age:cad.dur 1 3055.7 3067.7
<none>         3055.2 3069.2
- sex:cad.dur  1   3064.0  3076.0

Step:  AIC= 3067.27
tvdlm ~ sex + age + cad.dur + sex:cad.dur + age:cad.dur

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;none&gt;</td>
<td>3055.3</td>
</tr>
<tr>
<td>- age:cad.dur</td>
<td>1</td>
<td>3055.8</td>
</tr>
<tr>
<td>- sex:cad.dur</td>
<td>1</td>
<td>3064.1</td>
</tr>
</tbody>
</table>

Step:  AIC= 3065.79
tvdlm ~ sex + age + cad.dur + sex:cad.dur

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;none&gt;</td>
<td>3055.8</td>
</tr>
<tr>
<td>- sex:cad.dur</td>
<td>1</td>
<td>3066.8</td>
</tr>
<tr>
<td>- age</td>
<td>1</td>
<td>3105.9</td>
</tr>
</tbody>
</table>

Call:  glm(formula = tvdlm ~ sex + age + cad.dur + sex:cad.dur, family = binomial)

Coefficients:
<table>
<thead>
<tr>
<th>(Intercept)</th>
<th>sex</th>
<th>age</th>
<th>cad.dur</th>
<th>sex:cad.dur</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2.124102</td>
<td>-0.265944</td>
<td>0.034020</td>
<td>0.007418</td>
<td>-0.006221</td>
</tr>
</tbody>
</table>

Degrees of Freedom: 2331 Total (i.e. Null); 2327 Residual
Null Deviance: 3230
Residual Deviance: 3056 AIC: 3066
Starting with intercept only model

```r
> stepAIC(glm(tvdlm~-1+1,data=acath,family=binomial),scope=~sex*age*cad.dur)
Start:  AIC = 3232.49
  tvdlm ~ -1 + 1

Df  Deviance    AIC
+ cad.dur  1 3131.3 3135.3
+ age      1 3153.0 3157.0
+ sex      1 3217.0 3221.0
<none>           3230.5 3232.5

Step:  AIC = 3135.26
  tvdlm ~ cad.dur

Df  Deviance    AIC
  + age  1 3091.3 3097.3
  + sex  1 3117.9 3123.9
<none>           3131.3 3135.3
  - cad.dur  1 3230.5 3232.5
```
Step: AIC= 3097.32
      tvdlm ~ cad.dur + age

      Df  Deviance    AIC
      + sex  1  3066.8  3074.8
    <none>  3091.3  3097.3
      - age  1  3131.3  3135.3
      - cad.dur  1  3153.0  3157.0

Step: AIC= 3074.79
      tvdlm ~ cad.dur + age + sex

      Df  Deviance    AIC
    <none>  3066.8  3074.8
      - sex  1  3091.3  3097.3
      - age  1  3117.9  3123.9
      - cad.dur  1  3124.0  3130.0

Call: glm(formula = tvdlm ~ cad.dur + age + sex, family = binomial, data =

Coefficients:
(Intercept)    cad.dur    age    sex
-2.079777    0.005957    0.034330    -0.546153

Degrees of Freedom: 2331 Total (i.e. Null); 2328 Residual
Null Deviance: 3230
Residual Deviance: 3067    AIC: 3075

Which model do we prefer?
Prediction

An main interest of logistic regression is often prediction. Given that we estimate probabilities for individuals, how can we translate this into a predicted outcome? Two possibilities for prediction rules are:

1. Use 0.5 as a cutoff. That is if $\hat{\pi}$ for a new observation is greater than 0.5, its predicted outcome is $y = 1$. Otherwise, it’s $y = 0$. This approach is reasonable when

(a) it is equally likely in the population of interest that the outcomes 0 and 1 will occur, and
(b) the costs of incorrectly predicting 0 and 1 are approximately the same.
2. Find the best cutoff for the data set on which the multiple logistic regression model is based. Using this approach, we evaluate different cutoff values and for each cutoff value, calculate the proportion of observations that are incorrectly predicted. We would then select the cutoff value that minimizes the proportion of incorrectly predicted outcomes. This approach is reasonable when

(a) the data set is a random sample from the population of interest, and

(b) the costs of incorrectly predicting 0 and 1 are the same.
In the catheterization example,

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 \text{cad.dur}_i + \beta_2 \text{gender}_i,$$

if we use the cutoff of 0.5, we get the following results

```r
> table(fitted(glm11)>.5,acath$tvdlm)

   0   1
FALSE 937 674
TRUE  266 455
```

So, we misclassify people 40% of the time. Instead, let’s try finding a classification rule that minimizes misclassification in our data set.

```r
> for(p in seq(.35,.9,.05)){
+ t1=table(fitted(glm11)>p,acath$tvdlm)
+ (t1[2,1]+t1[1,2])/sum(t1)
+ 0.4030875
```
It looks like we can’t do much better than 40%.

What if we wanted to minimize missclassification for people with disease?

> for(p in seq(min(fitted(glm1)),.95,.05)){
+ t1=table(fitted(glm1)>p,acath$tvdlm)
+ cat(p, (t1[1,2])/sum(acath$tvdlm), (t1[2,1])/sum(1-acath$tvdlm), "\n")
+ }
0.329234 0.005314438 0.9925187
0.379234 0.08857396 0.8703242
0.429234 0.2604074 0.5652535
0.479234 0.5323295 0.2751455
0.529234 0.6589903 0.1704073
0.579234 0.765279 0.1080632
0.629234 0.8379097 0.06899418
0.679234 0.8990257 0.03740648
0.729234 0.9326838 0.01995012
0.779234 0.9619132 0.006650042
0.829234 0.9822852 0.004156276
0.879234 0.9911426 0.0008312552
0.929234 0.9991143 0
Quantifying predictive ability

Similar to the approach above we can plot the receiver operating characteristic (ROC) curve. This curve is a plot of 1-specificity against sensitivity.

We can plot this with a slight modification of the code above.

\[
p1 = \text{matrix}(0, nrow=13, ncol=3) \\
i = 1 \\
\text{for}(p \text{ in } \text{seq}(\text{min(fitted(glm1))}, .95, .05))\
\text{t1} = \text{table(fitted(glm1)} > p, \text{acath}$\text{tvdlm}$) \\
p1[i,] = c(p, (t1[2,2])/\text{sum(t1[,2]}), (t1[1,1])/\text{sum(t1[,1}])) \\
i = i + 1 \\
\text{plot}(1-p1[,3], p1[,2])
\]
The area under the ROC curve can give us insight into the predictive ability of the model. If it is equal to 0.5, the model can be thought of as predicting at random (an ROC curve with slope = 1). Values close to 1 indicate that the model has good predictive ability.

A similar measure is Somers’ $D_{xy}$ rank correlation between predicted probabilities and observed outcomes. It is given by

$$D_{xy} = 2(c - 0.5),$$

where $c$ is the area under the ROC curve. When $D_{xy} = 0$, the model is making random predictions. When $D_{xy} = 1$, the model discriminates perfectly.
We can get this $D_{xy}$ and $c$ value by using the `somers2()` function in the `Hmisc` library in R.

```r
> somers2(fitted(glmi1), acath$tvdlm)

              C      Dxy        n   Missing
0.6293747 0.2587493 2332.0000000 0.0000000
```

So, the area under the ROC curve is 0.629, and $D_{xy} = 0.26$. 
What if we add age to the model we’ve been looking at

\[
\text{logit}(\pi_i) = \beta_0 + \beta_1 \text{cad.dur}_i + \beta_2 \text{sex}_i + \beta_3 \text{age}_i
\]

|                | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | −2.0798  | 0.2575     | −8.08   | 0.0000   |
| cad.dur        | 0.0060   | 0.0008     | 7.22    | 0.0000   |
| sex            | −0.5462  | 0.1115     | −4.90   | 0.0000   |
| age            | 0.0343   | 0.0049     | 7.04    | 0.0000   |
\( c = 0.647 \) and \( D_{xy} = 0.295 \)