Some tools for assessing models - Part 1

Jarvis T. Chen (jarvis@hsph.harvard.edu)

20 October 2025

Harvard T.H. Chan School of Public Health

Learning Objectives

- Appreciate that **model assessment** is only possible when the goals of the modeling effort are well-defined and explicit
- Understand the concept of **nested models**, and understand F and likelihood ratio tests to evaluate the fit of a nested model relative to a larger model
- Understand the use of ”model fit statistics”, such as R^2, deviance, Akaike Information Criterion (AIC), and cross-validation measures, for assessing discrepancies between the model and data
- Appreciate the implications of model optimization for future prediction performance

A note on coding

I have included code and examples using the NHANES package in R. You'll want to make sure that you have the NHANES package and minimally dplyr loaded (or just load the tidyverse).

```r
if (!requireNamespace("NHANES", quietly = TRUE)) {
  install.packages("NHANES")
}

if (!requireNamespace("dplyr", quietly = TRUE)) {
  install.packages("dplyr")
}

library(NHANES)
library(dplyr)
```

The “true” model is elusive

- **Assumption of a known population model**: Up until now, we have worked under the assumption that the structure of the population model (i.e., the relationship between variables and the mean and variance functions) is known, apart from some unknown parameters.
    - This assumption has allowed us to make precise statements about how our estimators behave, particularly as the sample size grows large (asymptotically) or even in some finite-sample cases.
- **Reality of model uncertainty**: In most real-world applications, we do not know the ”true” model with certainty.
- **Model specification and its implications**: The models we fit are simplifications of reality. By choosing a specific model, we impose certain restrictions on the distribution of the data. These restrictions provide opportunities to test our ideas about the nature of the data generating process against the data.
- **Importance of a well-defined question**: Before building a model, the most crucial first step is to **pose a clear, well-defined research question**. What exactly do we want to learn or predict from the data?
    - Be explicit about the goals of the modeling process: Are we aiming to predict future outcomes, understand relationships between variables, or make causal inferences? This clarity guides the choice of model and its assumptions.

The goals of modeling

Example: What is the relationship between viral load in blood and risk of death in the next year among individuals with HIV?

- What is the goal of modeling this relationship?
    - Description of a set of data
    - Identify “risk factors”?
    - Predict the probability of an event?
    - Causal inference about modifiable exposures?
- Each of these goals has to be further specialized to determine the quantity we want to estimate
    - Parsimonious summary of a data distribution (data reduction)
    - A list of variables ranked by a measure of importance
    - A prediction model
    - An estimate of some causal quantity (e.g. average treatment effect of initiating treatment)

Model selection strategy needs to be tailored to the specific goals of modeling

- We cannot offer a strategy in the abstract
- That kind of approach does not recognize the importance of the goals of modeling (e.g. prediction, causal discovery, explanation, theory testing, policy making, etc.) in opting for a specific model selection approach
- Instead, we will talk about various approaches for evaluating the compatibility of models with the data
- These **model assessment** or **model fit assessment** methods are building blocks for refined/comprehensive model selection techniques
- In some cases, they can help us to pick among a small set of simple models or identify gross model inadequacies

Fit Statistics

- ”Fit statistic”: a function of the data that reflects the compatibility of the model with the data
- Explore relative overall performance of a (small, pre-specified) set of models
    - Often we will balance goodness-of-fit with model complexity (penalize overly complex models to avoid overfitting)
- Can identify gross model inadequacies where the model fails to capture key patterns in the data
- Fit statistics vary by model type
- Consider absolute vs. relative fit

Linear regression: R^2

- Partition total sum of squares (SST) into sum of squares explained (SSE) and sum of squares residuals (SSR):

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\text{SSE}} + \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{SSR}}$$

- Coefficient of determination: $R^2 = \frac{\text{SSE}}{\text{SST}} = \frac{\text{SST}-\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}$
- R^2 measures the proportion of the variability in the outcome explained by the model (i.e., by the predictor variables).
- **BUT**: R^2 always increases if additional covariates are added to a model!

A note on terminology:

- SSE (Sum of Squares Explained) Refers to the part of the total variability explained by the model, capturing the variance between the predicted values $\hat{y}_i$ and the mean $\bar{y}$. Note that sometimes it is referred to as the Sum of Squares Regression (SSR)!
- SSR (Sum of Squares Residuals) Refers to the part of the total variability that is not explained by the model, i.e., the variance between the observed values y_i and the predicted values $\hat{y}_i$.is sometimes referred to as the Sum of Squared Errors (SSE)!
- Be aware that some textbooks or fields may switch between these terms, but the underlying mathematical concepts remain the same.

Adjusted R^2

- Conceptually, R^2 looks like it is estimating the population R^2-: $\rho^2 = 1 - \frac{\sigma_\varepsilon^2}{\sigma_y^2}$, with estimators for σ_ε^2 (i.e., $\frac{\text{SSR}}{n-1}$) and σ_y^2 (i.e., $\frac{\text{SST}}{n-1}$) replacing their true values.
- However, we know that $\frac{\text{SSR}}{n-1}$ is a biased estimator for σ_ε^2
- What if we use the unbiased estimator: $\frac{\text{SSR}}{n-p-1}$, where $p+1$ is the number of parameters, including the intercept?
- This suggests the so-called **adjusted** R^2:

$$R^2_{\text{adj}} = 1 - \frac{\frac{\text{SSR}}{n-p-1}}{\frac{\text{SST}}{n-1}} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

- Some algebra shows that:

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

  This is often thought of as applying a penalty to R^2 as the number of parameters in the model increases. However, note that this does mean that R^2_{adj} can be negative.

- R^2_{adj} is generally considered useful only as a gross indicator of relative model fit.

Model fit for models estimated using likelihood methods

- For linear, logistic, and Poisson regression we built unified machnery based on maximum likelihood estimation
- Wouldn’t it be great if we also had a unified way to evaluate model fit and compare models ?

Example: logit(blood pressure > 130mmHg) $= \beta_0 + \beta_1(BMI)$

```
# Create a binary outcome variable: hbp (high blood pressure) defined as BPSysAve > 130
df_nhanes <- NHANES |>
  mutate(hbp = ifelse(BPSysAve > 130, 1, 0)) |>
  select(Age, Gender, SmokeNow, BMI, hbp) |>
  filter(!is.na(Age) & !is.na(Gender) & !is.na(SmokeNow) & !is.na(BMI) & !is.na(hbp))

# Fit the simpler logistic regression model (only Age and Gender as predictors)
model_basic <- glm(hbp ~ BMI, data = df_nhanes, family = binomial(link = "logit"))

summary(model_basic)
```

Example: logit(blood pressure > 130mmHg) $= \beta_0 + \beta_1(BMI)$

```
##
## Call:
## glm(formula = hbp ~ BMI, family = binomial(link = "logit"), data = df_nhanes)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.690394   0.186384  -9.069  < 2e-16 ***
## BMI          0.022868   0.006301   3.629 0.000284 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3551.3  on 3086  degrees of freedom
## Residual deviance: 3538.3  on 3085  degrees of freedom
## AIC: 3542.3
##
## Number of Fisher Scoring iterations: 4
```

- Generalized linear model
- Mean model specification: outcome variable and covariate
- Distribution assumption and link function

Nested Models

- One model is nested within another (larger) model if we can set some parameters to zero in the larger model and obtain the smaller model.
- For example:

  $$\log\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

  is nested within

  $$\log\left(\frac{\pi(X^*)}{1-\pi(X^*)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \beta_{p+1} X_{p+1} + \cdots + \beta_q X_q,$$

  where $q > p$.
- Which model above will fit a given dataset better?
- What would be a model with the poorest fit for a given dataset?
- What would be a model with the best fit for a given dataset?

Nested Models

- If the following two models are fitted to the same dataset, will the estimates of β_0 be the same in the two models?
    - Model 1:
    $$\log\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0$$
    - Model 2:
    $$\log\left(\frac{\pi(X^*)}{1-\pi(X^*)}\right) = \beta_0 + \beta_1 X_1$$
- If we calculate the value of the likelihood function at the MLE for Model 1 and also the value of the likelihood function at the MLE for Model 2, which value do you think will be larger?

Nested models and likelihood ratio tests

- **Conceptually:** For a given dataset, if a nested model fits the data well relative to the larger model, we would expect the ratio of likelihoods (calculated using the MLEs for each model) to be closer to one. This is compared to when the nested model fits the data poorly relative to the larger model.
- **Caveat:** Adding one or more covariates to a model will always improve the fit of the model to that given dataset.
- **Extreme Example:**
  - A model with as many parameters as observations would fit the data perfectly
  - For example, fitting a straight-line regression model (i.e., two parameters) to two data points will give a perfect fit.
  - On the other hand, a model with one parameter (such as a simple mean) would generally have a poorer fit.

Likelihood Ratio Test and Nested Models

- Nested Model: $\log\left(\frac{\pi(X)}{1-\pi(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
- Larger Model: $\log\left(\frac{\pi(X^*)}{1-\pi(X^*)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \beta_{p+1} X_{p+1} + \cdots + \beta_q X_q$

  where $q > p$

Likelihood Ratio Test (LRT):

To test the null hypothesis $H_0 : \beta_{p+1} = \cdots = \beta_q = 0$, use the test statistic:

$$R = -2\log\left(\frac{\text{Likelihood for nested model}}{\text{Likelihood for larger model}}\right)$$

Under H_0, $R \sim \chi^2_{q-p}$ asymptotically, where the likelihoods are calculated using the Maximum Likelihood Estimates (MLEs) for each model.

- Note: $R = -2\log(\text{likelihood for nested model}) + 2\log(\text{likelihood for larger model})$ and is easily calculated when software provides the log-likelihood for each model.
- Expectation of R under H_0: $\mathbb{E}(R \mid H_0) = q - p$ This reflects the caveat from the previous slide: adding variables to a model always gives a larger model that fits the specific dataset better.
- Application of LRTs: This approach is not specific to logistic regression and can be applied to other GLMs.

Example: high blood pressure in 2009-2012 NHANES

In this example, we will use logistic regression to predict the odds of having high blood pressure (defined as systolic blood pressure greater than 130 mmHg) using data from the NHANES 2009-2012 dataset. We will examine whether adding smoking status and BMI to a model that already includes age and gender improves the model's ability to predict high blood pressure.

The likelihood ratio test will help us determine whether the more complex model offers a statistically significant improvement in fit compared to the simpler model.

```r
# Create a binary outcome variable: hbp (high blood pressure) defined as BPSysAve > 130
df_nhanes <- NHANES |>
  mutate(hbp = ifelse(BPSysAve > 130, 1, 0)) |>
  select(Age, Gender, SmokeNow, BMI, hbp) |>
  filter(!is.na(Age) & !is.na(Gender) & !is.na(SmokeNow) & !is.na(BMI) & !is.na(hbp))

# Fit the simpler logistic regression model (only Age and Gender as predictors)
model_simple <- glm(hbp ~ Age + Gender, data = df_nhanes, family = binomial(link = "logit"))

# Fit the larger logistic regression model (Age, Gender, SmokeNow, and BMI as predictors)
model_large <- glm(hbp ~ Age + Gender + SmokeNow + BMI, data = df_nhanes, family = binomial(link = "logit"))

# Perform a Likelihood Ratio Test (LRT) using anova()
anova(model_simple, model_large, test = "Chisq")
```

Example: high blood pressure in 2009-2012 NHANES

```
## Analysis of Deviance Table
##
## Model 1: hbp ~ Age + Gender
## Model 2: hbp ~ Age + Gender + SmokeNow + BMI
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3084     3016.1
## 2      3082     3008.2  2   7.9175  0.01909 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F-Test for Nested Models in Linear Regression

- Nested Model (Smaller Model): $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$
- Larger Model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \beta_{p+1} X_{p+1} + \cdots + \beta_q X_q + \varepsilon$

where $q > p$.

F-Test

To test the null hypothesis $H_0 : \beta_{p+1} = \cdots = \beta_q = 0$, use the test statistic

$$F = \frac{\left(\frac{\text{SSR}_{\text{nested}} - \text{SSR}_{\text{larger}}}{q-p}\right)}{\left(\frac{\text{SSR}_{\text{larger}}}{n-q-1}\right)}$$

where:

- $\text{SSR}_{\text{nested}}$: Residual sum of squares for the nested model.
- $\text{SSR}_{\text{larger}}$: Residual sum of squares for the larger model.
- $q - p$: The number of additional parameters in the larger model.
- $n - q - 1$: Degrees of freedom for the larger model.

Under H_0, $F \sim F_{q-p, n-q-1}$: i.e. F-test statistic follows an F-distribution with $q - p$ and $n - q - 1$ degrees of freedom.

Example: systolic blood pressure in 2009-2012 NHANES

In this example, we will use linear regression to predict the average systolic blood pressure using data from the NHANES 2009-2012 dataset. We will examine whether adding smoking status and BMI to a model that already includes age and gender improves the model's ability to predict average systolic blood pressure.

Like the LRT, the F-test evaluates whether the more complex model offers a statistically significant improvement in fit compared to the simpler model.

```r
# Select relevant variables: Age, Gender, BPSysAve (systolic blood pressure), SmokeNow (smoking status), and BMI
df_nhanes <- NHANES |>
  select(Age, Gender, BPSysAve, SmokeNow, BMI) |>
  filter(!is.na(Age) & !is.na(Gender) & !is.na(BPSysAve) & !is.na(SmokeNow) & !is.na(BMI))

# Fit the simpler model (only Age and Gender as predictors)
model_simple <- lm(BPSysAve ~ Age + Gender, data = df_nhanes)

# Fit the larger model (Age, Gender, SmokeNow, and BMI as predictors)
model_large <- lm(BPSysAve ~ Age + Gender + SmokeNow + BMI, data = df_nhanes)

# Perform an F-test to compare the two models
anova(model_simple, model_large)
```

Example: systolic blood pressure in 2009-2012 NHANES

```
## Analysis of Variance Table
##
## Model 1: BPSysAve ~ Age + Gender
## Model 2: BPSysAve ~ Age + Gender + SmokeNow + BMI
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   3084 811811
## 2   3082 807995  2    3815.5 7.2769 0.0007033 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Relationship between the F-Test and Likelihood Ratio Test (LRT)

- For linear regression models with normally distributed errors, the F-test is related to the LRT because the residual sum of squares is proportional to the log-likelihood.
- Both tests evaluate whether adding additional parameters significantly improves the model fit, but the F-test is based on SSR, while the LRT is based on log-likelihoods.

Back to logistic regression: comparing our fitted model to a model that makes perfect predictions

- **Perfect Prediction Model in Logistic Regression:** In logistic regression, a model that makes perfect predictions would have $\hat{\pi}_i \in \{0, 1\}$, meaning the predicted probabilities exactly match the observed outcomes.

- We call this the **saturated model**, which is a model with as many parameters as there are data points

- The **deviance** for a fitted model is defined as:

$$D = -2 \log \left(\frac{\text{likelihood for the fitted model}}{\text{likelihood for the saturated model}} \right)$$

  This measures how much worse the fitted model's likelihood is compared to the saturated model's likelihood.

  - The deviance is a specific type of **likelihood ratio test statistic**. It compares the likelihood of our fitted model to the saturated model.
  - Conceptually, if a fitted model performs well relative to the saturated model, the likelihood ratio will be closer to one. As a result, the deviance D will be closer to zero.
  - In other words, a smaller deviance indicates a better fit, since the fitted model is closer to the saturated model, which represents perfect predictions.

Deviance for a Fitted Model

- The deviance for a fitted model is:

$$D = -2\log\left(\frac{\text{likelihood for the fitted model}}{\text{likelihood for the saturated model}}\right)$$

- **Interpretation of Deviance:** If the fitted model provides a poor fit to the data relative to the saturated model, then the likelihood ratio will be further from one, and D will be large.
  - The deviance captures the variability in the outcome that is not explained by the model.
  - In logistic regression (and other types of regression), the deviance plays a similar role to the *sum of squared residuals (SSR)* in standard linear regression.
  - Recall that in linear regression, the residual measures the difference between y_i (the observation, or what would be the prediction from the perfect/saturated model) and $\hat{y}_i$ (the prediction from the fitted model).
  - Like SSR in linear regression, the deviance has $n-(p+1)$ degrees of freedom, where n is the number of observations and p is the number of parameters in the model.
  - Hence, the deviance is sometimes referred to as the **residual deviance** (capturing the unexplained variability, similar to residual sum of squares in linear regression).

Example: high blood pressure in 2009-2012 NHANES

```
##
## Call:
## glm(formula = hbp ~ Age + Gender, family = binomial(link = "logit"),
##     data = df_nhanes)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.521811   0.186491 -24.247  < 2e-16 ***
## Age          0.061719   0.003007  20.522  < 2e-16 ***
## Gendermale   0.408238   0.092043   4.435 9.19e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3551.3  on 3086  degrees of freedom
## Residual deviance: 3016.1  on 3084  degrees of freedom
## AIC: 3022.1
##
## Number of Fisher Scoring iterations: 4
```

Example: high blood pressure in 2009-2012 NHANES

```
    Null deviance: 3551.3  on 3086  degrees of freedom
Residual deviance: 3016.1  on 3084  degrees of freedom
```

- $n - p - 1 = 3086 - 1 - 1 = 3084$
- If the model fits well compared to the perfect/saturated model, we expect the residual deviance to be close to d.f. based on chi-squared distribution
- For the above, $D = 3016.1$ on χ^2_{3084} would indicate significant lack of fit $(p < 0.05)$.

Deviance

- In ‘R‘, the ‘glm‘ function refers to the deviance of the fitted model as the **residual deviance**:

$$D_{\text{residual}} = -2\log\left(\frac{\text{likelihood for the fitted model}}{\text{likelihood for the saturated model}}\right)$$

- The degrees of freedom associated with the residual deviance is $n-(p+1)$, where n is the number of observations and p is the number of predictors.
- **Null Deviance:** The **null deviance** is calculated as:

$$D_{\text{null}} = -2\log\left(\frac{\text{likelihood for the null model}}{\text{likelihood for the saturated model}}\right)$$

  - The null model includes only a single parameter, the intercept.
  - The degrees of freedom for the null deviance is $n-1$, since there is only the intercept.

Deviance

- **Likelihood Ratio Test (LRT) for Model Fit:** We can construct a likelihood ratio test to compare the overall fit of the model with p parameters plus the intercept to the null model (intercept-only), i.e., testing the null hypothesis: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$
- The test statistic R is calculated as:

$$\begin{aligned} R &= -2\log\left(\frac{\text{likelihood for the null model}}{\text{likelihood for the fitted model}}\right) \\ &= -2\log\left(\frac{\text{likelihood for the null model}}{\text{likelihood for the saturated model}} \times \frac{\text{likelihood for the saturated model}}{\text{likelihood for the fitted model}}\right) \\ &= -2\log\left(\frac{\text{likelihood for the null model}}{\text{likelihood for the saturated model}}\right) + 2\log\left(\frac{\text{likelihood for the fitted model}}{\text{likelihood for the saturated model}}\right) \\ &= D_{\text{null}} - D_{\text{residual}} \end{aligned}$$

- R follows a chi-squared distribution with p degrees of freedom, where p is the number of predictors. $R \sim \chi^2_p$

Example: logit(blood pressure > 130mmHg) $= \beta_0 + \beta_1(BMI)$

```
##
## Call:
## glm(formula = hbp ~ BMI, family = binomial(link = "logit"), data = df_nhanes)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.690394   0.186384  -9.069  < 2e-16 ***
## BMI          0.022868   0.006301   3.629 0.000284 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3551.3  on 3086  degrees of freedom
## Residual deviance: 3538.3  on 3085  degrees of freedom
## AIC: 3542.3
##
## Number of Fisher Scoring iterations: 4
```

- $R = \text{(null deviance)} - \text{(residual deviance)} = 3551.3 - 3538.3 = 13 \sim \chi^2_1$
- `pchisq(chi_square_stat, df, lower.tail = FALSE)` gives a p-value of `0.000311491`
- This is a test of $H_0 : \beta_{\text{BMI}} = 0$. Where in the output is there another test of this H_0?

A note about deviance for logistic regression

- In logistic regression, if Y is coded as 0 and 1, we can show that the likelihood for the saturated model equals 1.
- The deviance D is calculated as:

$$
\begin{aligned}
D &= -2\log\left(\frac{\text{likelihood for the fitted model}}{\text{likelihood for the saturated model}}\right) \\
&= -2\log(\text{Likelihood for fitted model}) + 2\log(\text{Likelihood for saturated model}) \\
&= -2\log(\text{Likelihood for fitted model}) + 2\log(1) \\
&= -2\log(\text{Likelihood for fitted model})
\end{aligned}
$$

- Some software reports the value of D (deviance) instead of the log-likelihood value $\log(L)$.
  - This is useful to know when constructing likelihood ratio tests to compare two fitted models, one of which is nested within the other.
  - R, Stata, and SAS all report deviance. If you need the raw log-likelihood values instead of deviance, most software allows you to retrieve it via specific commands:
    - In R, use `logLik()` to extract the log-likelihood.
    - In Stata, use `lrtest` or `estat ic`.
    - In SAS, check the ”Fit Statistics” section for the log-likelihood.

Akaike Information Criterion (AIC)

- Basis in maximum likelihood theory: $AIC = -2\log(\hat{L}) + 2(p+1)$ where $\hat{L}$ is the maximum value of the likelihood function for the model, and p is the number of predictors in the model.
  - Prefer models with a lower AIC.
  - Models with an AIC that differs from the ”best” model by less than 2 are generally considered good.
  - A difference in AIC greater than 8 or 10 indicates a poor model relative to the best model.
- AIC can be used for any type of model fitted using maximum likelihood estimation, including logistic regression, linear regression, and generalized linear models.
  - For linear regression with a quantitative response variable, AIC can be written as:

$$AIC = n\log\left(\frac{\text{SSR}}{n}\right) + 2(p+1)$$

  where n is the sample size, SSR is the sum of squared residuals, and
  - p is the number of predictors.
  - Sometimes, AIC is written as: $AIC = n\log(\text{SSR}) + 2(p+1)$ (This version drops the term that only involves the sample size, which does not affect model comparison when using the same dataset).

Example: logit(blood pressure > 130mmHg) $= \beta_0 + \beta_1(BMI)$

```
##
## Call:
## glm(formula = hbp ~ BMI, family = binomial(link = "logit"), data = df_nhanes)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.690394   0.186384  -9.069  < 2e-16 ***
## BMI          0.022868   0.006301   3.629 0.000284 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3551.3  on 3086  degrees of freedom
## Residual deviance: 3538.3  on 3085  degrees of freedom
## AIC: 3542.3
##
## Number of Fisher Scoring iterations: 4
```

- For logistic regression with 0/1 coding for the outcome, $AIC =$ residual deviance $+ 2(p+1)$
- For this example, $p = 1$ and so $AIC = 3538.3 + 2(2) = 3542.3$
- Note that the AIC for the intercept only (”null”) model is $3551.3 + 2 = 3553$, so clearly the null model is worse than the model that includes BMI.

Cross-Validation

- One intuitive way to assess the relative performance of different models is to evaluate their predictive power on **out-of-sample** data, rather than relying on within-sample metrics like R^2. Assessing performance in this way ensures that the model generalizes well to new data, rather than just fitting the specific sample it was trained on.
  - **Training dataset**: This portion of the data is used to fit (or ”train”) the model.
  - **Validation dataset**: The remaining data is used to test the model’s predictions and evaluate its performance on data it hasn’t seen before.
- **Concept of Cross-Validation:** After training the model on one part of the data (the training dataset), we use it to predict outcomes on the validation dataset (the data not used for fitting). The predicted outcomes are then compared with the observed outcomes in the validation dataset to evaluate the model’s predictive power. This process provides an unbiased estimate of the model’s performance on new data.
- Models that are over-parameterized (overfitted) tend to have excellent performance (small residuals) on the training data because they capture even noise in the data. However, when applied to the validation data (data excluded from the fitting process), they often exhibit large prediction errors.

“Leave 1 Out” Cross-Validation Example

- Assume we have a sample of 100 observations and are considering three different models.
- **Step 1:** Regress y on x using candidate models 1, 2, and 3, excluding (i.e., leaving out) observation 1.
- **Step 2:** Predict the outcome for observation 1 for each of the three models.
- **Step 3:** Compute the square of the prediction error for each model:

  $$(y_1 - \hat{y}_{1,-1})^2$$

  where y_1 is the observed outcome and $\hat{y}_{1,-1}$ is the predicted outcome for observation 1 from the model excluding observation 1.
- **Repeat Steps 1-3:** Perform this process for all 100 observations by leaving out observations 2, 3, ..., 100 in turn.
- **Sum of Squared Prediction Errors (PRESS):** After repeating this process for all observations, sum the squared prediction errors for each model to calculate the predicted residual sum of squares (PRESS):

  $$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2$$

  for each model.
- **Select the Best Model:** Choose the model with the lowest PRESS as the ”best” model.
- Various other cross-validation techniques exist, such as “leave k out” or “exhaustive” cross-validation, which looks at all possible divisions of the sample into training and validation subsamples.

Computing the PRESS Statistic Using Leverage Values for Linear Regression

- For linear regression, we can compute PRESS without actually having to refit the model leaving out each observation by using the leverage values from the **hat matrix**.

- **PRESS residual for the i-th observation:**

$$e_{i,-i} = \frac{e_i}{1 - h_i}$$

  where e_i is the ordinary residual for observation i ($e_i = y_i - \hat{y}_i$), and h_i is the leverage for observation i, which is the diagonal element of the hat matrix $H = X(X^T X)^{-1} X^T$.

- **Formula for PRESS:**

$$\text{PRESS} = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_i} \right)^2$$

  This formula avoids having to re-fit the model for each observation.

- **Why this works:** The leverage values h_i measure the influence of each data point on its fitted value. By adjusting the residuals using the leverage, we account for the effect of leaving out each observation without having to re-fit the model multiple times.

Example

```r
# Select relevant variables and remove missing values
df_nhanes <- NHANES %>%
  select(Age, Gender, BPSysAve, SmokeNow, BMI) %>%
  filter(!is.na(Age) & !is.na(Gender) & !is.na(BPSysAve) & !is.na(SmokeNow) & !is.na(BMI))

# Fit the simpler model (only Age and Gender as predictors)
model_simple <- lm(BPSysAve ~ Age + Gender, data = df_nhanes)

# Fit the larger model (Age, Gender, SmokeNow, and BMI as predictors)
model_large <- lm(BPSysAve ~ Age + Gender + SmokeNow + BMI, data = df_nhanes)

# Function to compute PRESS statistic
compute_press <- function(model) {
  # Get the residuals from the model
    residuals <- resid(model)
  # Get the leverage values from the hat matrix
    hat_values <- hatvalues(model)
  # Compute PRESS residuals: e_i / (1 - h_i)
    press_residuals <- residuals / (1 - hat_values)
  # Compute PRESS: sum of squared PRESS residuals
    press_stat <- sum(press_residuals^2)
  return(press_stat)
}

press_simple <- compute_press(model_simple)
press_large <- compute_press(model_large)

cat("PRESS for the simpler model (Age + Gender):", press_simple, "\n")
cat("PRESS for the larger model (Age, Gender, SmokeNow, BMI):", press_large, "\n")
```

Example:

```
## PRESS for the simpler model (Age + Gender): 813456.1

## PRESS for the larger model (Age, Gender, SmokeNow, BMI): 810787.2
```

The larger model has the smaller (better) PRESS value.

Cross-validation limitations

- The main challenge with cross-validation is that it requires a lot of computational power, especially for large datasets and complex models.
- For p predictor variables, the number of possible model combinations is given by 2^p. For example, if $p = 2$, there are 4 possible model combinations, but if $p = 10$ there are 1024 possible combinations.
- As the number of predictors increases, the number of combinations grows exponentially, making it computationally expensive to explore all models.
- In large datasets, the number of possible validation subsamples becomes very large. Exploring all of them with Leave-One-Out Cross-Validation (LOO-CV) can be computationally prohibitive, potentially taking days or even longer, depending on the sample size and computing power.
- k-fold cross validation maybe more feasible since it splits the data into k roughly equal parts (or ”folds”). The model is fitted k times (one for each fold), which dramatically reduces the number of model fits compared to LOO-CV.
- Cross-validation is most practical when considering only a few specific models with moderately sized samples.

Fit statistics: summary

- There are a relatively large number of **fit statistics** that can be used in practice to assess the relative performance of models.
  - Most fit statistics involve a measure of **lack of fit**, such as the sum of squared residuals (SSR) or deviance. Many of these are penalized based on the number of parameters in the model to avoid overfitting.
- These can be a useful guide for identifying models that seem to fit the data well.
  - They allow us to compare models and assess their relative quality based on how well they explain the data while considering model complexity.
- Fit statistics can help us understand which covariates are consistently important across models, which are not important, and which (groups of) covariates might be interchangeable.
- **Challenge of Many Variables:** In practice, we often have a large number of variables and possible variable combinations, which makes it difficult and time-consuming to assess all potential models.

Model optimization and prediction

- Whenever we fit a model, we obtain parameter estimates that optimize the fit of the model to the dataset we have available.
- This is one kind of optimization, e.g. minimizing the sum of squares or maximizing the likelihood in order to obtain optimum parameter estimates for a given model
- Selecting covariates to include in a model adds another layer to model optimization. This involves balancing model complexity with predictive accuracy.
- **Question:** If we then use our selected model to make predictions for members of the population, how would the quality of those predictions generally compare to the quality of predictions made for the dataset used to select/fit the model?
- Consider the potential overfitting: the model may perform better on the training dataset (used to fit the model) but could perform worse when applied to new, unseen data due to overfitting or the inclusion of noise as predictive signals.

Guiding Principles for Model Building

- **Have a well-defined question and make the goals of modeling explicit:**
    - Consider conceptual or causal frameworks, and review the published literature.
    - Use intuition and common sense in model development.
- Give thought to the core set of covariates that are considered essential and should be included in any model, based on the scientific question or framework.
- Fit statistics (e.g., AIC, BIC) may be (moderately) useful for exploring possible models but should be viewed with caution.
- Useful for understanding which covariates are generally important, which are not important, and which (groups of) covariates may be interchangeable
- Cross-validation is an intuitive approach but can be computationally intensive, especially for large datasets.
- Stepwise and other automated variable selection algorithms should generally be avoided
- **Balancing art and science in model selection:** all criteria are just providing basic guidance for model specification (we are trying to make our intuitions more principled and therefore more ”scientific”)

Some helpful readings (not required)

Model-Building Strategies

Hosmer, Lemeshow and Sturdivant, *Applied Logistic Regression*, 3rd Edition, Chapter 4: Model-Building Strategies and Methods for Logistic Regression (although focused on logistic regression with some nice examples, ideas apply to statistical models in general; can skip/skim section 4.2.1 about using fractional polynomials though this is good material to know).

Underfitting and overfitting; some model fitting

Wooldridge, Introductory Econometrics, Chapter 3 and 6.3

Model selection

Draper and Smith (1998), *Applied Regression Analysis*, New York: Wiley. Chapter 15 (and Chapter 17 for ridge regression)

Cross-validation

Arlot S, Celisse A (2010) A survey of cross-validation procedures for model selection. *Stat Surv* 4: 40–79. doi: 10.1214/09-ss054

Assessing Model Fit in Logistic Regression

Assessing Model Fit in Logistic Regression: Hosmer, Lemeshow and Sturdivant. *Applied Logistic Regression*. Sections 5.1 and 5.2 Also Section 5.3 for those interested in going beyond the material on the notes regarding other diagnostics for assessing model fit in logistic regression.