BST 210 Lab: Week 8
Model Building & Goodness of Fit for Logistic Regression

Over the past two weeks, we've used logistic regression as our primary tool for estimating associations between and make predictions about a binary outcome Y and a set of predictors $X_1, \dots X_p$:

$$\text{logit}\big(P(Y=1)\big) = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p.$$

But there are many possible ways of modeling our outcome of interest—how and why did we settle on this one? As was the case for linear regression, we'd like to be able to formally compare different logistic regression models, and to decide which of our potential predictors are most appropriate/important to include in our model.

Our general process for going about model building is exactly the same as before! We:

1. Determine our objective for building the model
2. Select some metric that we'll use to assess which models are "best" (ex: AIC, BIC, covariate p-values)
3. Use subject matter knowledge and/or an automated procedure (such as forward selection, backward selection, stepwise selection, or best subsets selection) to arrive at a potential model
4. Validate the appropriateness/fit of our final model

We also want to keep the same guiding principles of parsimony and hierarchical well-formulation in mind.

However, as we saw in class this week, there are several additional wrinkles when it comes to logistic regression…

Maximum Likelihood Estimation

In linear regression, we fit our models by selecting the coefficient values that minimize the squared residual terms. In other words, we choose the parameter estimates that give us the smallest

$$SSE = \sum \epsilon_i^2 = \sum (y_i - \beta_0 - \beta_1 X_1 - \ldots - \beta_p X_p)^2.$$

This is why we are able to calculate/interpret:

- The sum of squares decomposition
- The Adjusted R^2
- The (root) MSE
- The F test statistic

In logistic regression, we instead estimate our coefficients so that we maximize the likelihood of our observed data—this approach is called **maximum likelihood estimation**.

Suppose that we have a data set with n individuals, and that for each of these individuals we observe a single binary outcome Y_i, and a single predictor X_i. The logistic regression relating X_i and Y_i is

$$\text{logit}\big(P(Y_i = 1)\big) = \beta_0 + \beta_1 X_i \implies P(Y_i = 1) = \frac{e^{\beta_0+\beta_1 X_i}}{1 + e^{\beta_0+\beta_1 X_i}}.$$

Given this particular form of the model, the likelihood of the observed data is given by

$$\begin{aligned}
\mathcal{L}(\beta_0, \beta_1) &= \prod_{i=1}^{n} P(Y_i = 1)^{y_i} \big(1 - P(Y_i = 1)\big)^{1-y_i} \\
&= \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_1 X_i}}{1 + e^{\beta_0+\beta_1 X_i}}\right)^{y_i} \left(1 - \frac{e^{\beta_0+\beta_1 X_i}}{1 + e^{\beta_0+\beta_1 X_i}}\right)^{1-y_i} \\
&= \prod_{i=1}^{n} \left(\frac{e^{\beta_0+\beta_1 X_i}}{1 + e^{\beta_0+\beta_1 X_i}}\right)^{y_i} \left(\frac{1}{1 + e^{\beta_0+\beta_1 X_i}}\right)^{1-y_i}.
\end{aligned}$$

When estimating β_0 and β_1, we choose the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ that make this likelihood—or equivalently the log likelihood, $\log(\mathcal{L})$—the largest.

Can we still use an F test to compare nested logistic regression models? If not, what other test(s) can we use instead?

Can we still use the Adjusted R^2 and MSE to compare non-nested models? If not, what metric(s) can we use instead?

An Example: Model Building in the GLOW Dataset

Let's return to the same data set that we saw last week: the Global Longitudinal Study of Osteoporosis in Women (GLOW). The data set is available on Canvas in the file `glow.csv`. Once again, our outcome of interest will be whether or not a study participant has a fracture within the first year of follow-up. Let's suppose that our objective is to build a prediction model for this probability of fracture, and that we have all of the variables in Table 1 at our disposal!

We'll start by reading the data set into SAS:

```
* Reading in the GLOW dataset;
proc import file='glow.csv' out=glow dbms=csv replace;
        getnames=yes;
run;
```

Table 1: Global Longitudinal Study of Osteoporosis in Women - Relevant Variables

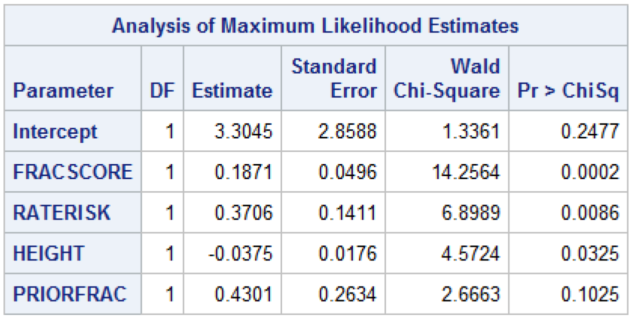
Variable	Description
SUB_ID	Subject identification code (numbered 1 through 500)
SITE_ID	Study site
PHY_ID	Physician ID code
PRIORFRAC	Indicator of history of prior fracture
AGE	Age at enrollment in the study
WEIGHT	Weight at enrollment in the study
HEIGHT	Height at enrollment in the study
BMI	BMI at enrollment in the study
PREMENO	Whether menopause occurred before age 45 (= 1) or after (= 0)
MOMFRAC	Whether the subject's mother had a hip fracture (= 1) or did not (= 0)
ARMASSIST	Whether arms are needed to stand from a chair (= 1) or not (= 0)
SMOKE	Whether a subject is a former or current smoker (= 1) or not (= 0)
RATERISK	Self-reported risk of fracture, recorded as 1 (less than others of the same age), 2 (same as others of the same age), or 3 (greater than others of the same age)
FRACSCORE	Composite fracture risk score
FRACTURE	Indicator for whether any fracture occurred in the first year of follow-up

While we could select and implement any of the automatic model building procedures we've talked about, let's focus on forward selection and backward elimination using a p-value criterion of 0.15. *How do these two model selection procedures work again?*

We can run forward selection with logistic regression in SAS much the same way we did for linear regression!

```
* Running a forward selection procedure;
proc logistic data=glow descending;
        model fracture = fracscore raterisk smoke armassist momfrac premeno
                bmi height age priorfrac / selection=forward slentry=0.15;
run;
```

After four steps, no remaining variables are significant at the $\alpha = 0.15$ entry level, and the the forward selection process stops:



Analysis of Maximum Likelihood Estimates

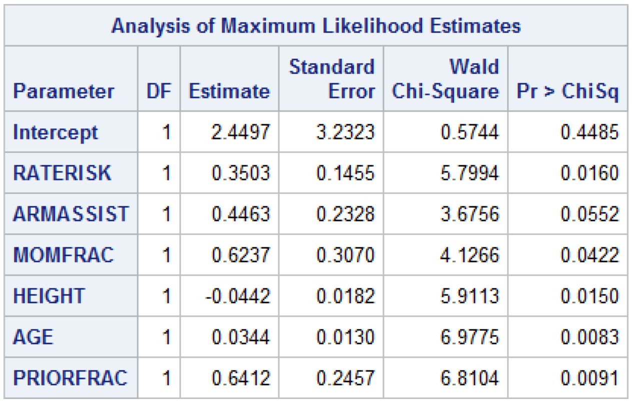
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	3.3045	2.8588	1.3361	0.2477
FRACSCORE	1	0.1871	0.0496	14.2564	0.0002
RATERISK	1	0.3706	0.1411	6.8989	0.0086
HEIGHT	1	-0.0375	0.0176	4.5724	0.0325
PRIORFRAC	1	0.4301	0.2634	2.6663	0.1025

So our final fitted model from forward selection is

$$\text{logit}(p) \;=\; 3.30 + 0.187 \cdot \texttt{fracscore} + 0.371 \cdot \texttt{raterisk} - 0.038 \cdot \texttt{height} + 0.430 \cdot \texttt{priorfrac}.$$

We can alternatively run a backward elimination procedure in SAS using the following code:

```
* Running a backward elimination procedure;
proc logistic data=glow descending;
        model fracture = frascore raterisk smoke armassist momfrac premeno
                bmi height age priorfrac / selection=backward slstay=0.15;
run;
```



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.4497	3.2323	0.5744	0.4485
RATERISK	1	0.3503	0.1455	5.7994	0.0160
ARMASSIST	1	0.4463	0.2328	3.6756	0.0552
MOMFRAC	1	0.6237	0.3070	4.1266	0.0422
HEIGHT	1	-0.0442	0.0182	5.9113	0.0150
AGE	1	0.0344	0.0130	6.9775	0.0083
PRIORFRAC	1	0.6412	0.2457	6.8104	0.0091

What is our final fitted logistic regression model from the backward elimination procedure?

Note that we arrived at two different "best" models for the probability of developing a fracture over the course of the study! *Are these two final models nested? If so, which is the full model and which is the reduced model?*

If they are not nested, how might we decide between them?

Assessing Goodness-of-Fit via Calibration

One possible goodness-of-fit metric that we discussed in lecture this past week is **calibration**. A logistic regression model is well-calibrated if the predicted probabilities we get from our regression model are reasonably close to the true probabilities in our observed data—in other words, that our model is a reasonable approximation of reality.

There are several key terms and ideas to keep in mind when we talk about calibration:

- **Covariate pattern:** a particular combination of covariate values
  - When we talk about calibration, we are typically concerned with the total number of covariate patterns in our observed data *for the covariates included in our model*
- **Saturated model:** a model that includes as many parameters as there are possible covariate patterns

Suppose that—when modeling the probability of a study participant having a fracture within the first year of follow-up—we had only included one predictor: the prior fracture indicator.

$$\text{logit}(p) = \beta_0 + \beta_1 \cdot \texttt{priorfrac}.$$

How many covariate patterns are there? Is this model saturated? And if not, what would the saturated model be?

Now suppose that we also add in the self-reported risk of fracture, represented as a categorical variable:

$$\text{logit}(p) = \beta_0 + \beta_1 \cdot \texttt{priorfrac} + \beta_2 \cdot I(\texttt{raterisk} = 1) + \beta_3 \cdot I(\texttt{raterisk} = 2).$$

How many covariate patterns are there now? Is this model saturated? And if not, what would the saturated model be?

Finally, suppose that we also included information on a participant's BMI at study entry. Then:

$$\text{logit}(p) = \beta_0 + \beta_1 \cdot \texttt{priorfrac} + \beta_2 \cdot I(\texttt{raterisk} = 1) + \beta_3 \cdot I(\texttt{raterisk} = 2) + \beta_4 \cdot \texttt{BMI}.$$

How many covariate patterns are there now? Is this model saturated? And if not, what would the saturated model be?

Generally speaking, what—if anything—can we say about the predicted probabilities/calibration of a saturated model?

Note that just because a particular model is saturated or well-calibrated, that does not necessarily mean that it includes all important predictors, or that it does a good job of describing variation in our outcome due to other covariates not in our model![1]

The Pearson Chi-Square Test

Suppose that we fit a logistic regression model with p predictors, and that in our data set there are J covariate patterns involving those predictors. For the j^{th} of those patterns, suppose that we have n_j individuals, O_j observed events, and $E_j = n_j \cdot \widehat{p}_j$ expected events (based on our model), and that we estimate the variability in the observed event counts by $V_j = n_j \cdot \widehat{p}_j(1-\widehat{p}_j)$. Then as long as the total number of covariate patterns J is small relative to our sample size, we can use the Pearson Chi-Square Test to assess the calibration of our model!

Hypothesis:	(H_0: the model has an acceptable fit) versus (H_1: the model does not fit well)
Test Statistic:	$\sum_{j=1}^{J} \frac{(O_j-E_j)^2}{V_j} \sim \chi^2_{J-(p+1)}$
Intuition:	How similar are our predicted probabilities/number of events to the truth?

Recall the two models that we arrived at using forward selection and backward elimination:

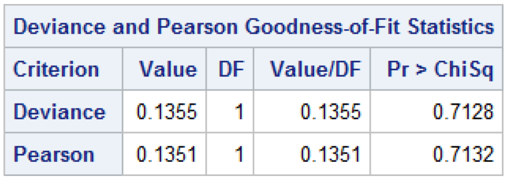
$$\begin{aligned}
\text{logit}(p) &= 3.30 + 0.187 \cdot \texttt{fracscore} + 0.371 \cdot \texttt{raterisk} - 0.038 \cdot \texttt{height} + 0.430 \cdot \texttt{priorfrac} \\
\text{logit}(p) &= 2.45 + 0.350 \cdot \texttt{raterisk} + 0.446 \cdot \texttt{armassist} + 0.624 \cdot \texttt{momfrac} - 0.044 \cdot \texttt{height} \\
&\quad + 0.0334 \cdot \texttt{age} + 0.641 \cdot \texttt{priorfrac}
\end{aligned}$$

Can we use the Pearson Chi-Square Test to assess either of their calibration? If so, what distribution will we compare the test statistic to? If not, why are we unable to perform the test?

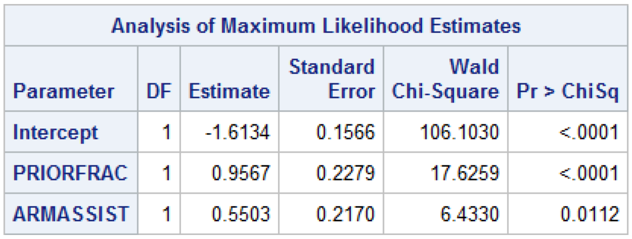
[1] To address this issue, some individuals define covariate patterns and saturated models differently. They consider all possible covariate patterns in the *entire* observed data set, including covariates that are not included in the model. SAS, Stata, and R assess calibration using our definition of covariate patterns/saturation, and in any sort of exam setting the two approaches will likely coincide.

For the purposes of seeing this test in action, let's consider a model that just includes `priorfrac` and `armassist`:

```
* Performing the Pearson Chi-Square test;
proc logistic data=glow descending;
        model fracture = priorfrac armassist / aggregate scale=none;
run;
```



Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	0.1355	1	0.1355	0.7128
Pearson	0.1351	1	0.1351	0.7132



Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.6134	0.1566	106.1030	<.0001
PRIORFRAC	1	0.9567	0.2279	17.6259	<.0001
ARMASSIST	1	0.5503	0.2170	6.4330	0.0112

What are our conclusions about the goodness-of-fit of the above model?

The Hosmer-Lemeshow Test

If we have a large number of covariate patterns—and in particular if we have a continuous covariate in our logistic regression model—we can't use the Pearson Chi-Square Test to assess goodness-of-fit. So we instead rely on a second test: the Hosmer-Lemeshow Test.

- Creates pseudo-covariate patterns by ranking the predicted probability for each individual, and then grouping observations into G groups of approximately equal sizes on the basis of these predicted probabilities
- Usually choose $G = 10$, corresponding to splitting the fitted probabilities into deciles

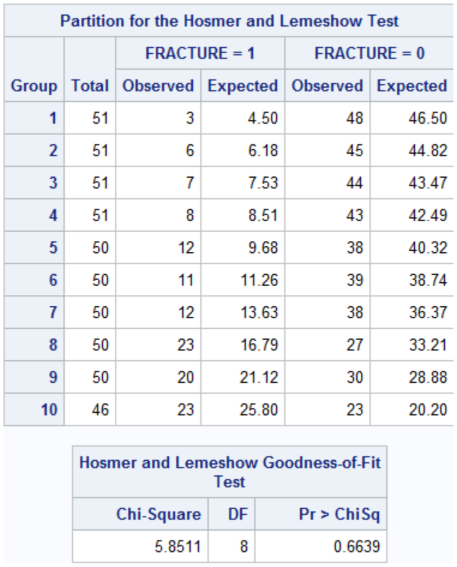
Hypothesis: (H_0: the model has an acceptable fit) versus (H_1: the model does not fit well)

Test Statistic: $\sum_{j=1}^{G} \frac{(O_j - E_j)^2}{n_j \bar{p}_j (1-\bar{p}_j)} \sim \chi^2_{G-2}$

Intuition: Similar to the Pearson Chi-Square Test with user-defined groups of observations

We can use the Hosmer-Lemeshow test to formally assess whether our forward selection and backward elimination models are well-calibrated:

```
* Performing the H-L Test for the forward selection model;
proc logistic data=glow descending;
        model fracture = fracscore raterisk height priorfrac /lackfit;
run;

* Performing the H-L Test for the backward elimination model;
proc logistic data=glow descending;
        model fracture = raterisk armassist momfrac height age priorfrac /lackfit;
run;
```



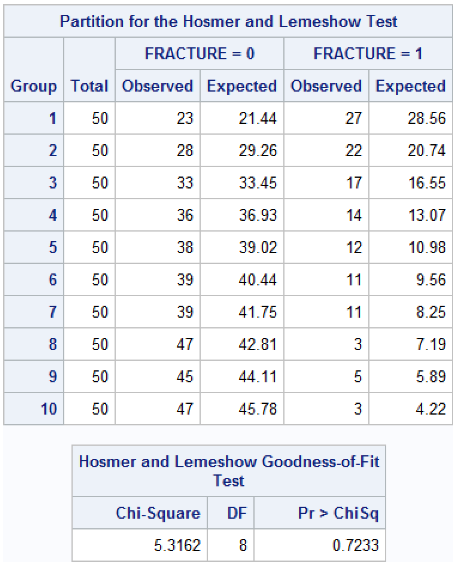
Partition for the Hosmer and Lemeshow Test

		FRACTURE = 1		FRACTURE = 0	
Group	Total	Observed	Expected	Observed	Expected
1	51	3	4.50	48	46.50
2	51	6	6.18	45	44.82
3	51	7	7.53	44	43.47
4	51	8	8.51	43	42.49
5	50	12	9.68	38	40.32
6	50	11	11.26	39	38.74
7	50	12	13.63	38	36.37
8	50	23	16.79	27	33.21
9	50	20	21.12	30	28.88
10	46	23	25.80	23	20.20

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
5.8511	8	0.6639

(a) Hosmer-Lemeshow results from the forward selection model.



Partition for the Hosmer and Lemeshow Test

		FRACTURE = 0		FRACTURE = 1	
Group	Total	Observed	Expected	Observed	Expected
1	50	23	21.44	27	28.56
2	50	28	29.26	22	20.74
3	50	33	33.45	17	16.55
4	50	36	36.93	14	13.07
5	50	38	39.02	12	10.98
6	50	39	40.44	11	9.56
7	50	39	41.75	11	8.25
8	50	47	42.81	3	7.19
9	50	45	44.11	5	5.89
10	50	47	45.78	3	4.22

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
5.3162	8	0.7233

(b) Hosmer-Lemeshow results from the backward elimination model.

What conclusions do we reach regarding the goodness-of-fit of the forward selected model? What about regarding the backward elimination model?

How useful is assessing goodness-of-fit via calibration when we want to compare and decide between models?

A note: usually we think of smaller p-values as being “better”, as they typically denote that some effect or relationship of interest is significant. However, in the case of goodness-of-fit, we *want* our p-values to be non-significant!

Assessing Goodness-of-Fit via Discrimination

Alternatively, we can assess a model's goodness-of-fit by looking at its **discrimination**—its ability to correctly separate out those individuals who actually had the event (i.e., those individuals with $Y = 1$) from those individuals who did not ($Y = 0$).

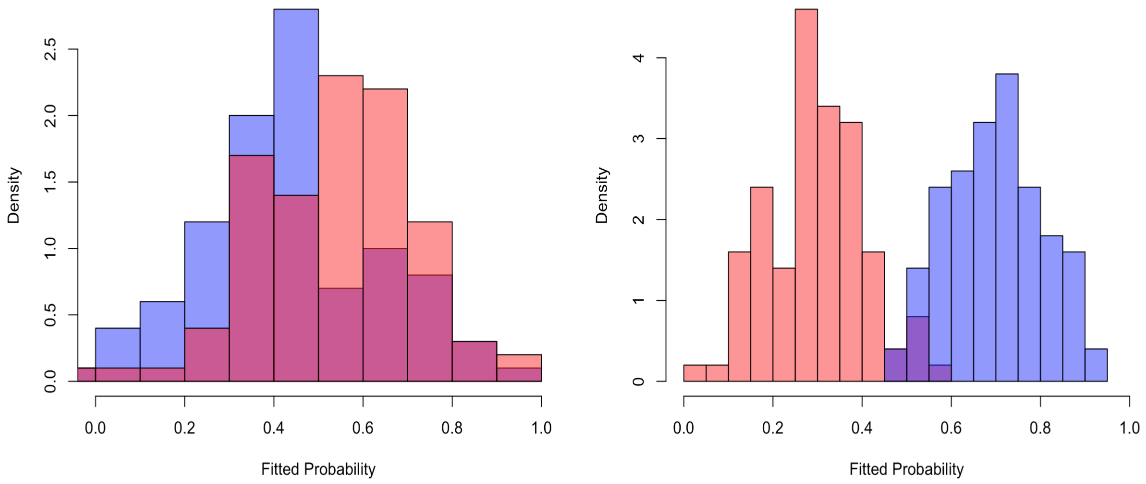


Figure 2: Fitted probabilities among those who are known to be cases (red) and known to be controls (blue). The model on the left has low discrimination, while the model on the right has high discrimination.

We often (eventually) want to use our fitted probabilities to actually predict whether or not a particular individual is a case (i.e. has $Y = 1$). To do this, we first need to choose some cut-off value, p_c, to use when classifying observations! We'll classify an individual as a case if their fitted probability of having an event, $\widehat{p}_i \geq p_c$, and we'll classify an individual as a control if their fitted probability of having an event, $\widehat{p}_i \leq p_c$.

- **Sensitivity:** the probability that we correctly classify someone as being a case, given that they truly are a case ($Y = 1$)
$$P\big(\,\widehat{p}_i \geq p_c \mid \text{the individual is a case}\big)$$
- **Specificity:** the probability that we correctly classify someone as a non-case, given that they truly are a non-case ($Y = 0$)
$$P\big(\,\widehat{p}_i < p_c \mid \text{the individual is a non-case}\big)$$

Note that—for any choice of p_c—there is a trade-off between the resulting sensitivity and specificity. *For example, what happens to our sensitivity and specificity if we choose $p_c = 0$?*

What happens if we choose $p_c = 1$?

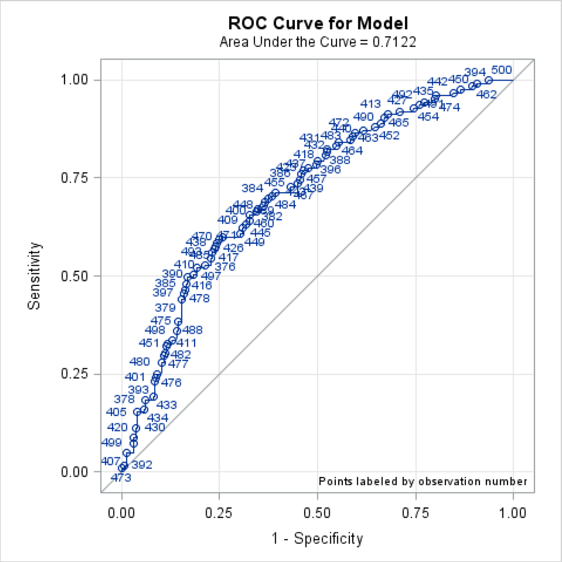
Models with higher discrimination make it easier for us to find a cut-point p_c that produces good results!

The ROC Curve

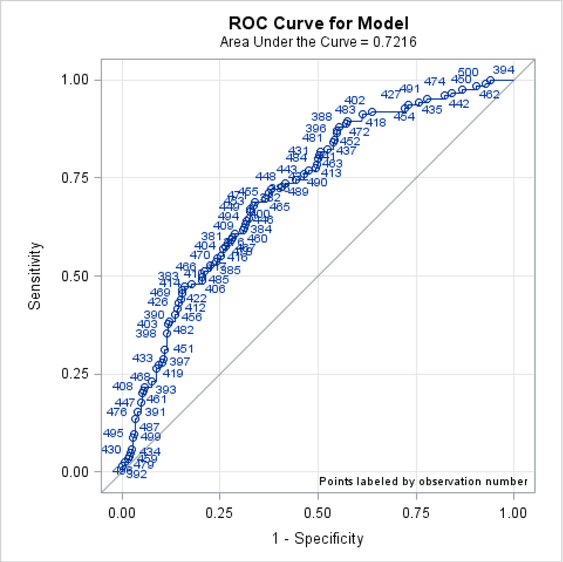
The ROC Curve (receiver-operator curve) helps us to visualize the trade-off between sensitivity and specificity, and to determine how well our logistic regression model discriminates between cases and controls. It plots our model's corresponding sensitivity and specificity for every possible choice of p_c.

Let's return to the GLOW data set, and plot the ROC curves for our two possible fracture models!

```
ods graphics on;

* Forward selection model;
proc logistic data=glow descending plots(only)=roc(id=obs);
        model fracture = fracscore raterisk height priorfrac
run;

* Backward elimination model;
proc logistic data=glow descending plots(only)=roc(id=obs);
        model fracture = raterisk armassist momfrac height age priorfrac;
run;

ods graphics off;
```



(a) ROC for forward selection model.



(b) ROC for backward elimination model.

Note that the area under the curve (**AUC**) is also sometimes known as the **c statistic**: it corresponds to the probability that a case (an individual with $Y = 1$) has a higher fitted probability than a non-case (an individual with $Y = 0$). *Do we prefer larger or smaller values of the AUC?*

Comparing the two ROC curves above, what do you notice? In particular, which model appears to provide a better fit to the data?

What would it mean if the ROC curve fell exactly on the y=x line? What if it fell below the y=x line?

Finally, based on all of the goodness-of-fit results above—as well as subject matter knowledge and the general principles of model building—which of the two fracture models do you prefer?

Midterm Exam Review

An Overview of Linear and Logistic Regression

Suppose that we have information on a continuous outcome Y_1 (say, systolic blood pressure), a binary outcome Y_2 (say, hypertension), and two predictors X_1 (stress levels) and X_2 (genetic background). We fit the following two models:

$$
\begin{aligned}
E[Y_1|X_1, X_2] &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 && (1) \\
\text{logit}\big(P(Y_2 = 1)\big) &= \beta_0 + \beta_1 X_1 + \beta_2 X_2. && (2)
\end{aligned}
$$

Compare and contrast the linear and logistic regression models on each of the following points:

What assumptions do we make in fitting these models?

How do we assess the appropriateness/goodness-of-fit of these models?

How can we incorporate possible non-linear effects into these models?

How do we get predicted values from these models?

How would we interpret the intercept β_0?

How would we interpret the slope for X_1, β_1?

How might we use these models to assess whether X_2 is a meaningful confounder of the association between X_1 and Y_1/Y_2? What additional models, if any, would we need to build to make this assessment?

How might we use these models to test whether X_1 is a significant predictor of Y_1/Y_2? What additional models, if any, would we need to build to make this assessment?

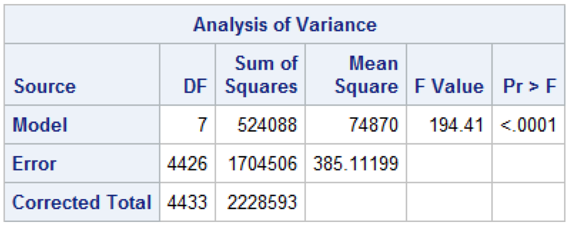
How would we use these models to test whether X_1 is an effect modifier of the association between X_2 and Y_1/Y_2? What additional models, if any, would we need to build to make this assessment?

A Deep Dive into Testing

Consider the following results from the linear regression of systolic blood pressure (measured in mmHg) on current smoking status, age, and categorical BMI (< 18.5, $18.5-25$, $25-30$, > 30), using 4434 observations:

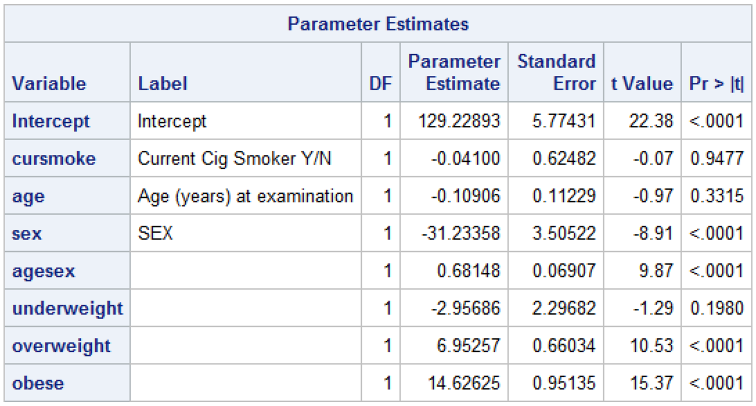
$$E[\texttt{sysbp}|X] = \beta_0 + \beta_1\texttt{cursmoke} + \beta_2\texttt{age} + \beta_3\texttt{sex} + \beta_4\texttt{age}\cdot\texttt{sex} + \beta_5\texttt{underweight} + \beta_6\texttt{overweight} + \beta_7\texttt{obese}.$$

The output is below:



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	524088	74870	194.41	<.0001
Error	4426	1704506	385.11199		
Corrected Total	4433	2228593			



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	129.22893	5.77431	22.38	<.0001
cursmoke	Current Cig Smoker Y/N	1	-0.04100	0.62482	-0.07	0.9477
age	Age (years) at examination	1	-0.10906	0.11229	-0.97	0.3315
sex	SEX	1	-31.23358	3.50522	-8.91	<.0001
agesex		1	0.68148	0.06907	9.87	<.0001
underweight		1	-2.95686	2.29682	-1.29	0.1980
overweight		1	6.95257	0.66034	10.53	<.0001
obese		1	14.62625	0.95135	15.37	<.0001

Is age significantly associated with systolic blood pressure among individuals with sex $= 0$*? How about among individuals with sex* $= 1$*? Construct and interpret a 95% Confidence Interval for each of these effects. Note that the covariance between* $\widehat{\beta}_2$ *and* $\widehat{\beta}_4$ *is -0.00736.*

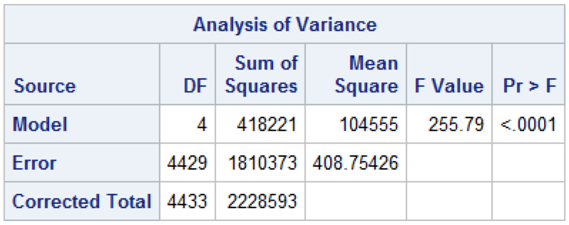
Suppose that we wish to test whether overweight individuals have mean systolic blood pressures that are 10 mmHg higher than individuals with BMI between 18.5 and 25. What are our null and alternative hypothesis? What is our test statistic, and what distribution do we compare it to?

Based just on the output above, do we have enough information to tell whether categorical BMI is a significant predictor of systolic blood pressure?

Suppose that we also fit the linear regression model that omits all categorical BMI terms:

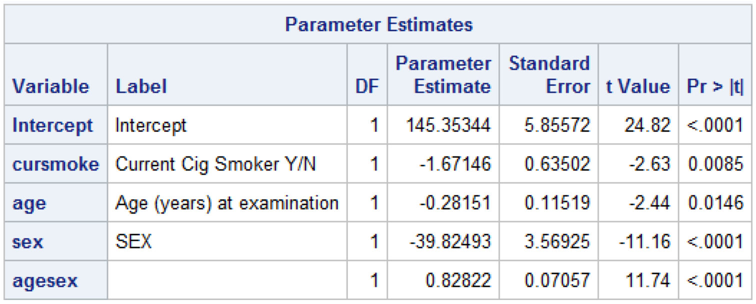
$$E[\texttt{sysbp}|X] = \beta_0 + \beta_1\texttt{cursmoke} + \beta_2\texttt{age} + \beta_3\texttt{sex} + \beta_4\texttt{age}\cdot\texttt{sex}.$$

The output is below:



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	418221	104555	255.79	<.0001
Error	4429	1810373	408.75426		
Corrected Total	4433	2228593			



Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	145.35344	5.85572	24.82	<.0001
cursmoke	Current Cig Smoker Y/N	1	-1.67146	0.63502	-2.63	0.0085
age	Age (years) at examination	1	-0.28151	0.11519	-2.44	0.0146
sex	SEX	1	-39.82493	3.56925	-11.16	<.0001
agesex		1	0.82822	0.07057	11.74	<.0001

Using this additional information, conduct the formal hypothesis test for assessing whether or not categorical BMI is a significant predictor of systolic blood pressure.

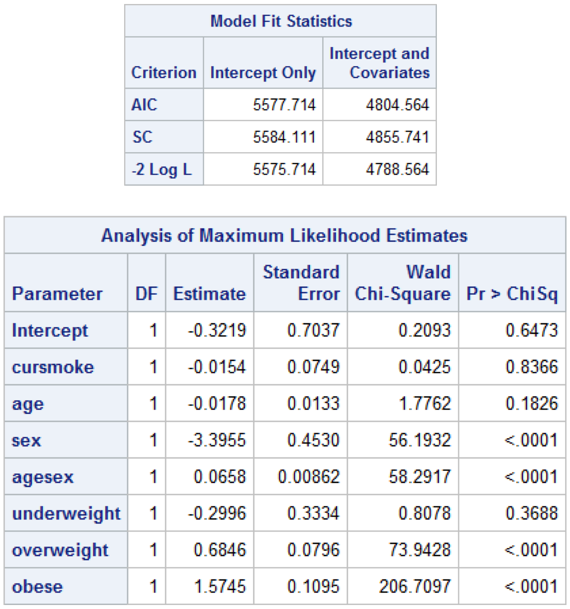
Suppose that we now fit the same two models, but that this time we take hypertension (dichotomized as "yes"/"no") as our outcome of interest:

$$\text{logit}(p) = \beta_0 + \beta_1\texttt{cursmoke} + \beta_2\texttt{age} + \beta_3\texttt{sex} + \beta_4\texttt{age}\cdot\texttt{sex} + \beta_5\texttt{underweight} + \beta_6\texttt{overweight} + \beta_7\texttt{obese}$$

and

$$\text{logit}(p) = \beta_0 + \beta_1\texttt{cursmoke} + \beta_2\texttt{age} + \beta_3\texttt{sex} + \beta_4\texttt{age}\cdot\texttt{sex}.$$

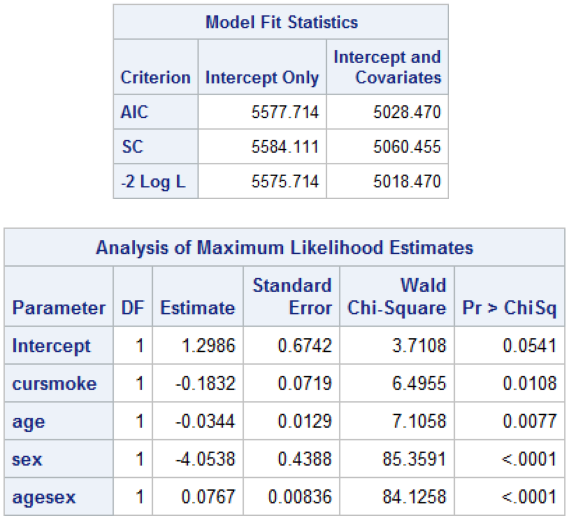
The output is on the next page.



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	5577.714	4804.564
SC	5584.111	4855.741
-2 Log L	5575.714	4788.564

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.3219	0.7037	0.2093	0.6473
cursmoke	1	-0.0154	0.0749	0.0425	0.8366
age	1	-0.0178	0.0133	1.7762	0.1826
sex	1	-3.3955	0.4530	56.1932	<.0001
agesex	1	0.0658	0.00862	58.2917	<.0001
underweight	1	-0.2996	0.3334	0.8078	0.3688
overweight	1	0.6846	0.0796	73.9428	<.0001
obese	1	1.5745	0.1095	206.7097	<.0001

(a) Output from the full model



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	5577.714	5028.470
SC	5584.111	5060.455
-2 Log L	5575.714	5018.470

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.2986	0.6742	3.7108	0.0541
cursmoke	1	-0.1832	0.0719	6.4955	0.0108
age	1	-0.0344	0.0129	7.1058	0.0077
sex	1	-4.0538	0.4388	85.3591	<.0001
agesex	1	0.0767	0.00836	84.1258	<.0001

(b) Output from the reduced model

In the full model (on the right), is age significantly associated with hypertensive status among individuals with sex $= 0$*? How about among individuals with sex* $= 1$*? Construct and interpret a 95% Confidence Interval for each of these effects. Note that the covariance between* $\widehat{\beta}_2$ *and* $\widehat{\beta}_4$ *is -0.000109.*

Assess whether categorical BMI is a significant predictor of hypertensive status.