BST 210 Lab: Week 6
Concluding Linear Regression & Motivating Logistic Regression

Multiple linear regression models allow us to assess the relationship between a continuous outcome Y and a set of predictors $X_1, \ldots, X_p$. One (now very familiar!) way of writing this mathematically is

$$E[Y|X_1, \ldots, X_p] = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p.$$

For the first a few weeks of class, we've mainly focused on how to construct this model and select the predictors $X_1, \ldots, X_p$.

1 Hypothesis Testing and Linear Models

But we also want to be able to use this model to draw statistical conclusions about our predictors $X_1, \ldots, X_p$ and their relationship with Y! There are several main types of hypotheses that we may want to test:

- $H_0 : \beta_j = 0$ for a specific predictor X_j
- $H_0 : \beta_j = \beta_k$ for two predictors X_j and X_k
- $H_0 : \beta_i = \beta_j = \beta_k = 0$ for a collection of predictors $\{X_i, X_j, X_k\}$
- $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ for all predictors $\{X_i, X_j, X_p\}$

$\boxed{H_0 : \beta_j = 0}$

Here, we want to assess whether a statistically significant relationship between X_j and Y exists. In other words, we want to answer the question: is X_j a statistically significant predictor of our outcome, Y?

Formal Hypothesis:	$H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$
Test Statistic:	$t = \frac{\hat{\beta}_j - 0}{\text{s.e.}(\hat{\beta}_j)} \sim t_{n-(p+1)}$
Confidence Interval:	$\hat{\beta}_j \pm t_{n-(p+1),1-\frac{\alpha}{2}} \cdot \text{s.e.}(\hat{\beta}_j)$

$\boxed{H_0 : \beta_j = \beta_k}$

Here, we want to assess whether a linear combination of the βs is significantly different than zero.

Formal Hypothesis:	$\big(H_0 : \beta_j = \beta_k$ versus $H_1 : \beta_j \neq \beta_k\big)$ or $\big(H_0 : \beta_j - \beta_k = 0$ versus $H_1 : \beta_j - \beta_k \neq 0\big)$
Test Statistic:	$t = \frac{(\hat{\beta}_j - \hat{\beta}_k) - 0}{\text{s.e.}(\hat{\beta}_j - \hat{\beta}_k)} \sim t_{n-(p+1)}$
Confidence Interval:	$(\hat{\beta}_j - \hat{\beta}_k) \pm t_{n-(p+1),1-\frac{\alpha}{2}} \cdot \text{s.e.}(\hat{\beta}_j - \hat{\beta}_k)$

This type of hypothesis test is also particularly useful in the setting where X_k is an interaction term, and where $\beta_j + \beta_k$ is the slope for the association between X_j and Y within a particular level of an effect modifier!

Formal Hypothesis:	$H_0 : \beta_j + \beta_k = 0$ versus $H_1 : \beta_j + \beta_k \neq 0$
Test Statistic:	$t = \frac{(\hat{\beta}_j + \hat{\beta}_k) - 0}{\text{s.e.}(\hat{\beta}_j + \hat{\beta}_k)} \sim t_{n-(p+1)}$
Confidence Interval:	$(\hat{\beta}_j + \hat{\beta}_k) \pm t_{n-(p+1),1-\frac{\alpha}{2}} \cdot \text{s.e.}(\hat{\beta}_j + \hat{\beta}_k)$

$\boxed{H_0 : \beta_i = \beta_j = \beta_k = 0}$

Here, we want to determine whether a particular subset of covariates—here X_i, X_j, and X_k—contributes significantly to our model. So we want to test whether, after accounting for all other covariates, X_i, X_j and X_k collectively explain a significant proportion of the remaining variability in Y.

We can equivalently view the test of the null hypothesis $H_0 : \beta_i = \beta_j = \beta_k = 0$ as comparing the fit of the **reduced model** without the covariates X_i, X_j, and X_k,

$$E[Y|X] = \beta_0 + \beta_1 \cdot X_1 + \ldots + 0 \cdot X_i + \ldots + 0 \cdot X_j + \ldots + 0 \cdot X_k + \ldots + X_p,$$

to the fit of the **full model**

$$E[Y|X] = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_p \cdot X_p.$$

So we can summarize this test by:

Formal Hypothesis:	$H_0 : \beta_i = \beta_j = \beta_k = 0$ versus H_1 : at least one of β_i, β_j and β_k is not 0
or	
Formal Hypothesis:	$\big(H_0$: the reduced model is sufficient$\big)$ versus $\big(H_1$: the full model is preferred$\big)$
Test Statistic:	$F = \frac{(SSE_{reduced} - SSE_{full})/3}{SSE_{full}/(n-(p+1))} \sim F_{3,n-p-1}$

where $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is the error sum of squares, this is also called RSS (residual sum of squares), but is different from SSR (regression sum of squares). In this lab, we will use SSE throughout.

We can fairly easily generalize this to a subset of the covariates of size r. *What would the test statistic and its distribution be in that case?*

If we want to test whether a subset of r covariates contributes significantly to our model, then our *reduced model* is the model of $E[Y|X]$ excluding our r covariates, and our *full model* is the model of $E[Y|X]$ with all covariates included. Then our test statistic is

$$F = \frac{(SSE_{reduced} - SSE_{full})/r}{SSE_{full}/(n-(p+1))} \sim F_{r,n-p-1}.$$

In fact, we can use this kind of F-test (also known as an ANOVA) to compare any two **nested** models. *What would the test statistic and its distribution be in that case?*

When we say that models are **nested**, we simply mean that we can arrive at our smaller, reduced model by placing some sort of constraints on our larger, full model. In other words, we can think of the full model as being the smaller model "plus something else". Suppose that we have p covariates in the full model and q covariates in the reduced model (so $p > q$). Then our test statistic is

$$F = \frac{(SSE_{reduced} - SSE_{full})/(p-q)}{SSE_{full}/(n-p-1)} \sim F_{p-q,n-p-1}$$

$\boxed{H_0 : \beta_1 = \ldots = \beta_p = 0}$

Here, we want to determine whether the mean model, $E[Y|X] = \beta_0$, is alone sufficient to explain the variability in our outcome Y.

Formal Hypothesis:	$H_0 : \beta_1 = \ldots = \beta_p = 0$ versus H_1 : at least one of $\beta_1, \ldots, \beta_p$ is not 0
Test Statistic:	$F = \frac{(SSE_{reduced} - SSE_{full})/p}{SSE_{full}/(n-(p+1))} = \frac{(SST - SSE_{full})/p}{SSE_{full}/(n-(p+1))} \sim F_{p,n-p-1}$

2 Hypothesis Testing: Example

For this example, we'll once again use a subset of the Framingham Heart Study, found in the file `framingham.dta`. A full summary of all the covariates in the dataset is given in Table 1 below. In this example, we'll be using total cholesterol (`totchol`) as our outcome of interest.

Let's suppose that we arrived at the model

$$E[\texttt{totchol}|X] = \beta_0 + \beta_1 \cdot \texttt{sex} + \beta_2 \cdot \texttt{age} + \beta_3 \cdot \texttt{diabp} + \beta_4 \cdot \texttt{cigpday} + \beta_5 \cdot \texttt{bmi}. \quad (1)$$

and would like to now perform inference on it. We'll start by fitting this model in Stata!

Table 1: Framingham Heart Study - Relevant Variables

Variable	Description
totchol	Serum Total Cholesterol (mg/dL)
sex	Participant sex - 1 = Men, 2 = Women
age	Age at exam (years)
sysbp	Systolic Blood Pressure (mean of last two of three measurements) (mmHg)
diabp	Diastolic Blood Pressure (mean of last two of three measurements) (mmHg)
cursmoke	Current cigarette smoking - 1 = Yes, 0 = No
cigpday	Number of cigarettes smoked each day
bmi	Body Mass Index, weight in kilograms/height meters squared
diabetes	Diabetic according to criteria of first exam treated or casual glucose of 200 mg/dL or more
prevhyp	Prevalent Hypertensive. Subject was defined as hypertensive if treated or, if second exam at which mean systolic was >=140 mmHg or mean Diastolic >=90 mmHg
prevchd	Prevalent Coronary Heart Disease

```
* Reading in the dataset
cd "your_path_here"
use "framingham.dta", clear

* Removing all missing observations
foreach v of var * {
 drop if mi(`v')
}

* Fitting the regression model
regress totchol sex age diabp cigpday bmi
```

Source	SS	df	MS	Number of obs	=	2,223
				F(5, 2217)	=	52.95
Model	471898.722	5	94379.7445	Prob > F	=	0.0000
Residual	3951303.33	2,217	1782.27485	R-squared	=	0.1067
				Adj R-squared	=	0.1047
Total	4423202.06	2,222	1990.63999	Root MSE	=	42.217

totchol	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	1.322259	.1075758	12.29	0.000	1.111299	1.533219
sex	8.612305	1.97406	4.36	0.000	4.741105	12.48351
diabp	.381843	.0821498	4.65	0.000	.2207443	.5429417
cigpday	.1805195	.0833422	2.17	0.030	.0170825	.3439565
bmi	.5199803	.2029924	2.56	0.010	.1219052	.9180554
_cons	108.9981	8.579065	12.71	0.000	92.17427	125.822

Let's first test to see whether or not the mean model alone (intercept only) is sufficient to explain the variability in total cholesterol. *What are our null and alternative hypotheses?*

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \quad vs \quad H_1$: at least one of the coefficients is non-zero

```
* Performing an F-test (testing all betas equal to 0)
test sex age diabp cigpday bmi
```

```
( 1)  sex = 0
( 2)  age = 0
( 3)  diabp = 0
( 4)  cigpday = 0
( 5)  bmi = 0

       F(  5,  2217) =   52.95
            Prob > F =    0.0000
```

Based on the output above, we reject the null hypothesis that the mean model alone is sufficient to explain the observed variability in total cholesterol ($p < 0.0001$). As such, we conclude that at least one of age, sex, diastolic blood pressure, number of cigarettes smoked per day, and BMI is a significant predictor of total cholesterol. [In Stata, the results of this F-test is always provided as part of the regression output!]

Suppose we want to look at one of those covariates, BMI—and its relationship with total cholesterol—a little more closely. *Using the regression output given above, conduct the t-test*

$$H_0 : \beta_5 = 0 \quad vs \quad H_1 : \beta_5 \neq 0.$$

What do you conclude, and how would you report this conclusion? Also report the 95% confidence interval for this effect.

We reject the null hypothesis that $\beta_5 = 0$, and conclude that BMI is significantly associated with mean total cholesterol, after adjusting for age, sex, diastolic blood pressure, and number of cigarettes smoked per day ($p = 0.01$). With 95% confidence, a one unit change in BMI is associated with between a 0.122 and 0.918 unit increase in mean total cholesterol, holding constant age, sex, diastolic blood pressure, and number of cigarettes smocked per day. Although Stata gives us this interval, we could also find it by calculating

$$0.520 \pm t_{2223-6,0.975} \cdot 0.203 \approx (0.122, 0.918).$$

Another way that we might go about testing whether BMI is a significant predictor of total cholesterol is through an F-test (ANOVA) in which we compare the full model including BMI to the reduced model without BMI. More specifically, we say that the full model is the one fit above,

$$E[\texttt{totchol}|X] = \beta_0 + \beta_1 \cdot \texttt{sex} + \beta_2 \cdot \texttt{age} + \beta_3 \cdot \texttt{diabp} + \beta_4 \cdot \texttt{cigpday} + \beta_5 \cdot \texttt{bmi},$$

while the reduced model is the one without BMI,

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

We can then directly compare these two models in Stata:

```
* Performing an F-test (testing beta_bmi equal to 0)
test bmi
```

```
 ( 1)  bmi = 0

       F(  1,  2217) =    6.56
            Prob > F =    0.0105
```

What do you notice, if anything, about the results of this test as compared to the results of the t-test for $\beta_5 = 0$?

The p-value that we get from the F-test is exactly the same as that for the t-test (the slight difference in reported p-values is just a rounding issue)! We also notice that the test statistic for the F-test, $F = 6.56$, is the square of the test statistic for the t-test: $T^2 = 2.56^2 = 6.56$.

In fact, it will always be the case that, if $T \sim t_{n-p-1}$, then $T^2 \sim F_{1,n-p-1}$. So the t-test for $\beta_j = 0$ will always give exactly the same conclusions as the F-test comparing the (full) model including X_j to the (reduced) model without X_j!

Finally, suppose that we want to determine whether or not we should add a spline of BMI to our model. *How can we test for whether linear BMI alone (without an additional flexible regression term) sufficiently captures the association between BMI and total cholesterol?*

This question should look familiar—we answered a very similar question as part of Homework 3! In short, we can use an F-test to formally test whether linear BMI alone sufficiently captures our association of interest, given that we already know the linear BMI model is nested in the spline BMI model.
For the purposes of performing the F-test, we'll consider the reduced model to be the model with just linear bmi, and the full model to be the model with both linear BMI and a spline. We can then test whether the coefficients corresponding to the spline terms of BMI are significantly different from 0; if these coefficients aren't significantly different from 0, we can conclude that linear BMI alone is sufficient.

```
* Making a cubic spline of bmi with 4 knots
mkspline bmis=bmi, cubic nknots(5)

* Fitting the full model (including spline term)
regress totchol age sex diabp cigpday bmi bmis1 bmis2 bmis3 bmis4
```

```
* Peforming an F-test (testing whether the slopes of the spline terms equal 0)
test bmis1 bmis2 bmis3 bmis4
```

```
 test bmis1 bmis2 bmis3 bmis4

( 1)  o.bmis1 = 0
( 2)  bmis2 = 0
( 3)  bmis3 = 0
( 4)  bmis4 = 0
      Constraint 1 dropped

      F(  3,  2214) =    9.80
           Prob > F =    0.0000
```

What are our conclusions?

We reject the null hypothesis that each of the coefficients for the spline term is equal to 0, and so conclude that at least one must be significantly different from 0 ($p < 0.0001$). As such, we conclude that the flexible spline term of BMI explains some significant proportion of the variability in total cholesterol after accounting for age, sex, diastolic blood pressure, number of cigarettes smoked per day, and linear BMI—so linear BMI alone is not sufficient.

An extra note about hypothesis testing: be aware there is no formal statistical test for the existence of confounding, so you have to check the causal definition with subject matter knowledge and the "10% change in coefficients" is just a rule of thumb. We do have formal statistcal tests for effect modification.

3 Relaxing the Assumptions for Linear Regression

In fitting and performing inference with a linear regression model, we make four assumptions:

1. **L**inearity
2. **I**ndependence
3. **N**ormality of the residuals
4. **E**qual variance (homoskedasticity)

However, in practice, it's rare that all four of these assumptions hold. So that leaves us with the question: if these assumptions are violated, what happens to our OLS estimates, and is there anything we can do to "fix" them?

If our assumption of linearity is incorrect, that means that our model is misspecified, and $E[Y|X_1, \ldots, X_p] \neq \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_2 \cdot X_p$. *In this scenario, what happens to our OLS estimates and standard errors? Is there anything we can do?*

If our model is misspecified, we're out of luck :(

Neither our coefficients nor our standard errors are estimating the right thing, and there really isn't much we can do about it.

Yet there are cases where we can get consistent estimators (the estimated values converges to the true parameters asymptotically).

However, all hope is not lost! *Under what conditions is our OLS estimate $\hat{\beta}$ still unbiased for the true coefficient β?*

Actually, as long as our model is correctly specified (i.e. as long as the linearity assumption holds) and our errors (residuals) have mean 0, our estimates $\hat{\beta}$ will be unbiased for the true β! This holds regardless of whether or not we have independence, normality of the residuals, or constant variance.

Suppose our model is correctly specified, but that at least one of the other three assumptions is violated. Are our estimates of the standard errors unbiased? If not, what can we do?

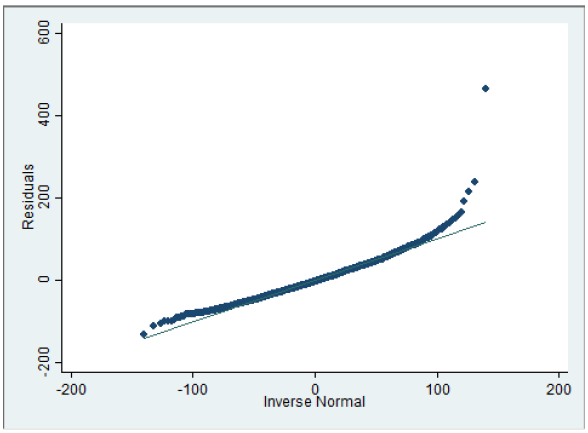
No, our standard error estimates will be biased if the independence and equal variances assumptions are violated. However, there are methods that exist to correct this bias! One simple way is to use the robust standard error instead.

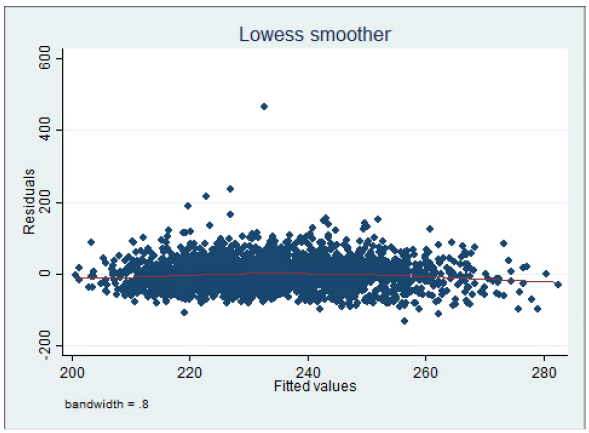
Let's go back to our Framingham Heart data set, and look at some diagnostic plots for regression model (1).

```
* Running Model (1) again
regress totchol age sex diabp cigpday bmi

* Storing the residuals and fitted values
predict lmresid, residuals
predict lmfit

* Creating residual plots for Model (1): qqplot and Lowess curve
qnorm lmresid
lowess lmresid lmfit
```





What do these plots suggest about our LINE assumptions?

Things don't look great... From the quantile-quantile plot, we can see a pretty strong departure from normality at both tails.

Robust standard errors are usually the most helpful when you have unequal variances, but you can use it even when the equal variances assumption holds (it is still valid, but you may lose some efficiency). Let's run a regression with robust standard errors here and compare it to the model using ordinary standard errors.

```
* Running Model (1) again, this time with a robust variance estimator
regress totchol age sex diabp cigpday bmi, vce(robust)
```

```
Linear regression                               Number of obs     =      2,223
                                                F(5, 2217)        =      53.10
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1067
                                                Root MSE          =     42.217

------------------------------------------------------------------------------
             |               Robust
    totchol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   8.612305   2.007778     4.29   0.000     4.674984    12.54963
         age |   1.322259   .1061529    12.46   0.000      1.11409    1.530429
       diabp |    .381843   .0810519     4.71   0.000     .2228974    .5407886
     cigpday |   .1805195   .0793498     2.27   0.023     .0249117    .3361272
         bmi |   .5199803   .2114899     2.46   0.014     .1052413    .9347193
       _cons |   108.9981   8.600798    12.67   0.000     92.13165    125.8646
------------------------------------------------------------------------------
```

How does this new model with robust standard errors compare to the original fit of Model (1)?

As we would expect, the coefficient estimates are exactly the same as they were in the standard regression model! The standard error estimates (and the confidence intervals) have been adjusted slightly, but in this case there have been no substantial corrections, and we reach the same conclusions from the t-tests of all of our coefficients.

Also note that we no longer get any information about the sum of squares decomposition!

What are some the "trade-offs" of using a more robust variance estimator?

Although robust variance estimates help protect against the violation of our LINE assumptions, we have to sacrifice some things. As mentioned before, we may lose some efficiency if the equal variances assumption actually holds. Also, the robust variance estimators are large sample variances, meaning that our t-tests and F-tests no longer hold exactly—we need to be concerned about this if the sample size is small. We also have the drawback that such measures as adjusted-R^2, AIC, BIC, etc. no longer have nice interpretations as before.

---

4 Binary Outcomes & Logistic Regression

Often times in public health, we're interested in estimating and modeling some sort of binary ("yes"/"no") outcome, rather than a continuous response. For example, we might be interested in

- Whether a tumor does or does not metastasize
- Whether a patient does or does not have a heart attack
- Whether a child does or does not have asthma

We can think of this binary outcome Y as taking on two distinct values: $Y = 1$ if the event of interest occurs, and $Y = 0$ otherwise. Then we would like to be able to make statements about $E[Y] = P(Y = 1) = p$.

Equivalently, we may also be interested in saying something about the *odds* of Y occurring, where the odds of Y are defined as $P(Y = 1)/P(Y = 0) = p/(1 - p)$.

4.1 Contingency Table Example

Consider the data in the contingency table below, which were collected as part of a study that investigated the effect of parental smoking status (X) on the smoking habits of students in Arizona (Y). The data are taken from the textbook *Categorical Data Analysis*, and can be found on Canvas in the file `smoker.dta` (Agresti 1990).

	At Least One Parent Smokes	Neither Parent Smokes	Total
Student Smokes	816	188	1004
Student Does Not Smoke	3203	1168	4371
Total	4019	1356	5375

What is $p_1 = P(Y = 1|X = 1)$*? What about* $p_0 = P(Y = 1|X = 0)$*?*

$p_1 : \frac{816}{4019} \approx 0.203$ & $p_0 : \frac{188}{1356} \approx 0.139$

One measure of effect that we can use to examine the relationship between parental smoking status and student smoking habits is the risk difference: $p_1 - p_0$*. Calculate and interpret this quantity.*

$p_1 - p_0 : 0.203 - 0.139 = 0.064$
We estimate that the risk of a student smoking is 0.064 greater among the students for whom at least one parent smokes compared to those for whom neither parent smokes.

Another effect measure we discussed in class was the risk ratio. Calculate and interpret the risk ratio for this data.

$\frac{p_1}{p_0} : \frac{0.203}{0.139} = 1.46$
We estimate that the risk of a student smoking is 1.46 times greater (equivalently 46% higher) among the students for whom at least one parent smokes compared to those for whom neither parent smokes.

The last measure of association—and the one that we will use most frequently in this class—is the odds ratio. Calculate and interpret this value.

Odds when $X = 1$: $\frac{p_1}{1-p_1} = \frac{0.203}{1-0.203} \approx 0.255$ & Odds when $X = 0$: $\frac{p_0}{1-p_0} = \frac{0.139}{1-0.139} \approx 0.161$

$OR: \frac{p_1}{1-p_1} \cdot \frac{1-p_0}{p_0} = \frac{816 \cdot 1168}{188 \cdot 3203} \approx 1.58$
We estimate that the odds of a student smoking is 1.58 times greater (58% greater) among the students for whom at least one parent smokes compared to those for whom neither parent smokes.

Construct and report a 95% confidence interval for the odds ratio you found above. Does this confidence interval contain 1? What does that suggest to you?

$Var\big(\ln(OR)\big) \approx \frac{1}{816} + \frac{1}{3203} + \frac{1}{188} + \frac{1}{1168} \approx 0.0077$
95% CI for $\ln(OR)$: $\ln(1.58) \pm 1.96 \cdot \sqrt{0.0077} \approx (0.285, 0.629)$

95% CI for OR: $(e^{0.285}, e^{0.629}) \approx (1.33, 1.88)$
With 95% confidence, the odds of a student smoking is estimated to be between 1.33 and 1.88 times greater among the students for whom at least one parent smokes compared to those students for whom neither parent smokes. Notice that this confidence interval excludes 1, which leads us to believe that the association between parental smoking habits and student smoking habits might be statistically significant.

We can recreate the contingency table found above in Stata by using the following commands:

```
* Reading in the .dta file
use "smoker.dta", clear

* Creating a contingency table for the relationship between student smoking and
* parental smoking
cs ssmoke psmoke [fweight=freq]
```

```
. cs ssmoke psmoke [fweight=freq], or woolf

                |      psmoke
                |   Exposed   Unexposed      Total
----------------+------------------------------------
          Cases |       816         188       1004
       Noncases |      3203        1164       4367
----------------+------------------------------------
          Total |      4019        1352       5371
                |
           Risk |   .2030356    .1390533   .1869298
                |
                |      Point estimate    [95% Conf. Interval]
----------------+------------------------------------
Risk difference |       .0639823         .0417378    .0862269
     Risk ratio |       1.460128         1.261661    1.689816
Attr. frac. ex. |       .3151286          .207394    .4082195
Attr. frac. pop |       .2561205
     Odds ratio |       1.577351         1.327879    1.873692 (Woolf)
----------------+------------------------------------
                                chi2(1) =    27.25  Pr>chi2 = 0.0000
```

Simple Logistic Regression

In simple logistic regression, we are now interested in modeling a *binary* outcome Y as a function of some predictor variable, X, where X can be either continuous or categorical.

Why is it that we can't use the simple linear regression model $p = E[Y|X] = \beta_0 + \beta_1 \cdot X$ like we did when Y was continuous?

Two main reasons: 1) probabilities are constrained to be between 0 and 1, and the standard linear regression framework does not impose/respect this restriction; 2) the variance of a binary outcome Y with success probability p depends on p—$Var(Y) = p(1-p)$—so it most certainly violates our assumption of constant

variance.

As such, the model we fit for logistic regression is instead in terms of the log odds of Y:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot X \quad \Longrightarrow \quad p = \frac{\exp\{\beta_0 + \beta_1 \cdot X\}}{1 + \exp\{\beta_0 + \beta_1 \cdot X\}}.$$

What is the intercept term, β_0, estimating? What about e^{β_0}?

β_0 estimates the log odds of Y in the population where $X = 0$
e^{β_0} estimates the odds of Y in the population where $X = 0$

For the purposes of this question, let's assume that X is also a binary variable. What is β_1? How do we interpret e^{β_1}?

β_1 estimates the log odds ratio of Y comparing the population where $X = 1$ and the population where $X = 0$.

e^{β_1} estimates the odds ratio of Y comparing the population where $X = 1$ and the population where $X = 0$.