Logistic Regression and Likelihood

EPI 204

Quantitative Epidemiology III

Statistical Models

Logistic Regression with Raw Data

- Most times the data are in the form of individual cases with the covariates and resulting binary classification variable as a 0/1 variable or two-level factor. It is convenient not to have to tabulate
- Also, if any of the covariates is continuous, categorization is not possible without discretizing the variable, which is often not a good idea.
- In the hypertension example, each of the eight categories had subjects with exactly the same values of the predictors. This is often not the case.

juul(ISwR) R Documentation

Juul's IGF data
Description
The juul data frame has 1339 rows and 6 columns. It contains a reference sample of the distribution of insulin-like growth factor (IGF-1), one observation per subject in various ages with the bulk of the data collected in connection with school physical examinations.

Format
This data frame contains the following columns:

age: a numeric vector (years).
menarche: a numeric vector. Has menarche occurred (code 1: no, 2: yes)?
sex: a numeric vector (1: boy, 2: girl).
igf1: a numeric vector. Insulin-like growth factor (μg/l).
tanner: a numeric vector. Codes 1–5: Stages of puberty a.m. Tanner.
testvol: a numeric vector. Testicular volume (ml).

Source
Original data.

Tanner Score

- The Tanner score is a measure of physical maturation based on secondary sex characteristics such as body hair, breast development, and genital development (Marshall and Tanner 1969, 1970).
- It is technically an *ordinal* variable with values 1, 2, 3, 4, 5 in order.
- Ordinal variables can be treated as such with specialized software.
- Another possibility is to treat them as linear as an approximation.
- We will use the Tanner score as a categorical variable with five levels.

```
> library(ISwR)
> data(juul)
> juul1 <- subset(juul,age > 8 & age < 20 & complete.cases(menarche))
```

Girls between 8 and 20 with non-missing menarche variable.

```
> summary(juul1)
      age            menarche          sex          igf1            tanner      
 Min.   : 8.03   Min.   :1.000   Min.   :2   Min.   : 95.0   Min.   : 1.000  
 1st Qu.:10.62   1st Qu.:1.000   1st Qu.:2   1st Qu.:280.5   1st Qu.: 1.000  
 Median :13.17   Median :2.000   Median :2   Median :409.0   Median : 4.000  
 Mean   :13.44   Mean   :1.507   Mean   :2   Mean   :414.1   Mean   : 3.307  
 3rd Qu.:16.48   3rd Qu.:2.000   3rd Qu.:2   3rd Qu.:514.0   3rd Qu.: 5.000  
 Max.   :19.75   Max.   :2.000   Max.   :2   Max.   :914.0   Max.   : 5.000  
                                             NA's   :108.0   NA's   :83.000  
    testvol      
 Min.   : NA     
 1st Qu.: NA     
 Median : NA     
 Mean   :NaN     
 3rd Qu.: NA     
 Max.   : NA     
 NA's   :519     
```

```
> juul1$menarche <- factor(juul1$menarche,labels=c("No","Yes"))
> juul1$tanner <- factor(juul1$tanner)
> attach(juul1)
> summary(glm(menarche ~ age,binomial))

Call:
glm(formula = menarche ~ age, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.32759  -0.18998   0.01253   0.12132   2.45922

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -20.0132     2.0284  -9.867   <2e-16 ***
age           1.5173     0.1544   9.829   <2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 719.39  on 518  degrees of freedom
Residual deviance: 200.66  on 517  degrees of freedom
AIC: 204.66
```

```
> summary(glm(menarche ~ age+tanner,binomial))

Call:
glm(formula = menarche ~ age + tanner, family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.56180  -0.12461   0.02475   0.08055   2.86120

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.7758     2.7630  -4.986 6.17e-07 ***
age           0.8603     0.2311   3.723 0.000197 ***
tanner2      -0.5211     1.4846  -0.351 0.725609
tanner3       0.8264     1.2377   0.668 0.504313
tanner4       2.5645     1.2172   2.107 0.035132 *
tanner5       5.1897     1.4140   3.670 0.000242 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 604.2  on 435  degrees of freedom
Residual deviance: 106.6  on 430  degrees of freedom
AIC: 118.6
```

```
> anova(glm(menarche ~ age+tanner,binomial),test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: menarche

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     435     604.19
age     1   442.31       434     161.88 3.396e-98
tanner  4    55.28       430     106.60 2.835e-11

> drop1(glm(menarche ~ age+tanner,binomial),test="Chisq")
Single term deletions

Model:
menarche ~ age + tanner
       Df Deviance     AIC     LRT   Pr(Chi)
<none>     106.599 118.599
age     1  124.500 134.500  17.901 2.327e-05 ***
tanner  4  161.881 165.881  55.282 2.835e-11 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1
```

SAS Version of the Analysis

```
*NA's in data file changed to .*
data juul ;
  infile '/folders/myfolders/juul.txt' firstobs=2;
  input obs $ age menarche sex igf1 tanner $ testvol;
  drop igf1 testvol;
  if age <= 8 then delete;
  if age >= 20 then delete;
  if missing(menarche) then delete;
 run;

 proc print data=juul( obs=10);
 run;

 proc logistic data=juul;
    class tanner (ref="1" param=ref);
    model menarche(desc) = age tanner;
run;
```

More on Odds Ratios

- Each coefficient except the intercept is an estimate of the log odds ratio between two conditions.
- If a factor has two levels (say “Yes” and “No”), then the two conditions are for individuals at “Yes” and individuals at “No” with other variables and factors held constant.
- If there are interaction terms, then one must specify the levels of the other predictors, often at the average.
- If a factor has more than two levels, then the coefficients compare a level with a baseline level and other comparisons have to be derived.
- For a continuous variable, the coefficient is the log odds ratio for a unit change in the variable.

More on Odds Ratios

The logistic model is

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

The LHS is the log odds. If x_i changes from 0 to 1, then the log odds changes by β_i. Since a difference of logs is the log of the ratio, β_i is the log of the odds ratio for a unit change in x_i.

We can view β_0 as the baseline log odds. This is meaningful only for the specific population at hand. If we have a case control study with 50% cases and 50% controls, and the population has 4% cases, then clearly the baseline risk for the case control study has no relevance to the population.

```
Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.7758     2.7630  -4.986 6.17e-07 ***
age           0.8603     0.2311   3.723 0.000197 ***
tanner2      -0.5211     1.4846  -0.351 0.725609
tanner3       0.8264     1.2377   0.668 0.504313
tanner4       2.5645     1.2172   2.107 0.035132 *
tanner5       5.1897     1.4140   3.670 0.000242 ***
```

Log odds ratio for one year increase in age is 0.8603
Odds ratio is exp(0.8603) = 2.364
Log odds ratio for a two year increase in age is (2)(0.8603) = 1.7206
Odds ratio is exp(1.7206) = 5.588
All these holding tanner score constant

Log odds ratio for tanner 4 vs. tanner 1 is 2.5645
Odds ratio is exp(2.5645) = 12.994
Log odds ratio for tanner 4 vs. tanner 3 is 2.5645 – 0.8264 = 1.7381
Odds ratio is exp(1.7381) = 5.687
All these holding age constant

Of course, age and Tanner score are correlated so the “holding constant” is a numerical calculation only.

Likelihood

- The likelihood is the pdf of the data thought of as a function of the parameters for data already observed.
- Maximum likelihood (ML) is an established method of estimating the parameters in a data analysis problem, though it sometimes may fail and often needs some alteration.
- The MLE of a Gaussian mean is the sample mean. But the MLE of the variance is the sum of squares of errors divided by n (not $n - 1$).
- In practice, we use the variance estimator with divisor $n - 1$ which is a small variant.

The binomial distribution has probability mass function

$$f(x \mid n, p) = P(X = x \mid n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

After we observe the data, x and n are known. Estimate p.

The likelihood is a function of p given x and n.

$$f(p \mid n, x) = \binom{n}{x} p^x (1-p)^{n-x}$$

This is maximized over p for fixed n and x when this is maximized:

$g(p) = x \ln(p) + (n-x) \ln(1-p)$ (log likelihood omitting first term).

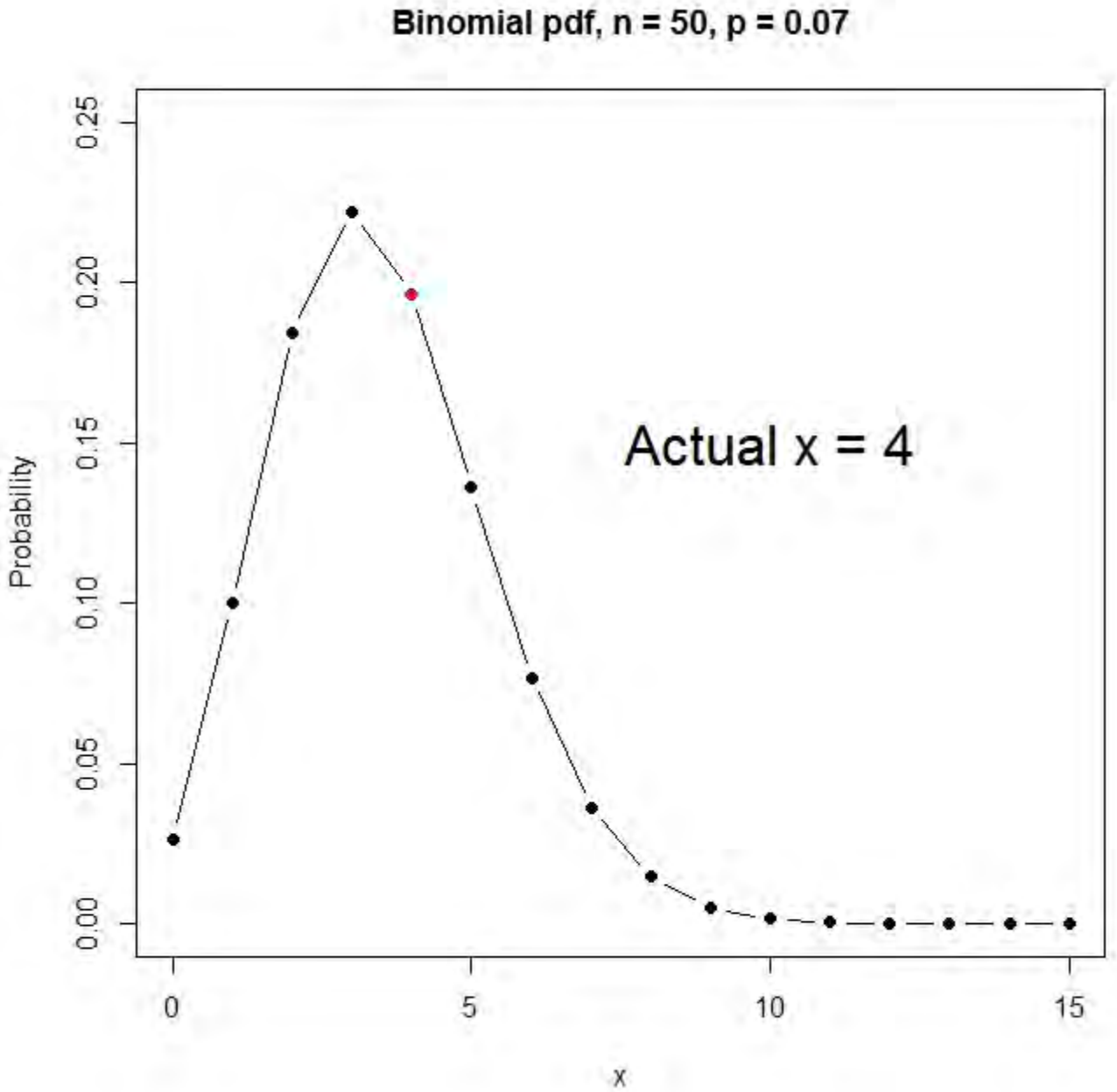
$g'(p) = x/p - (n-x)/(1-p) = 0$

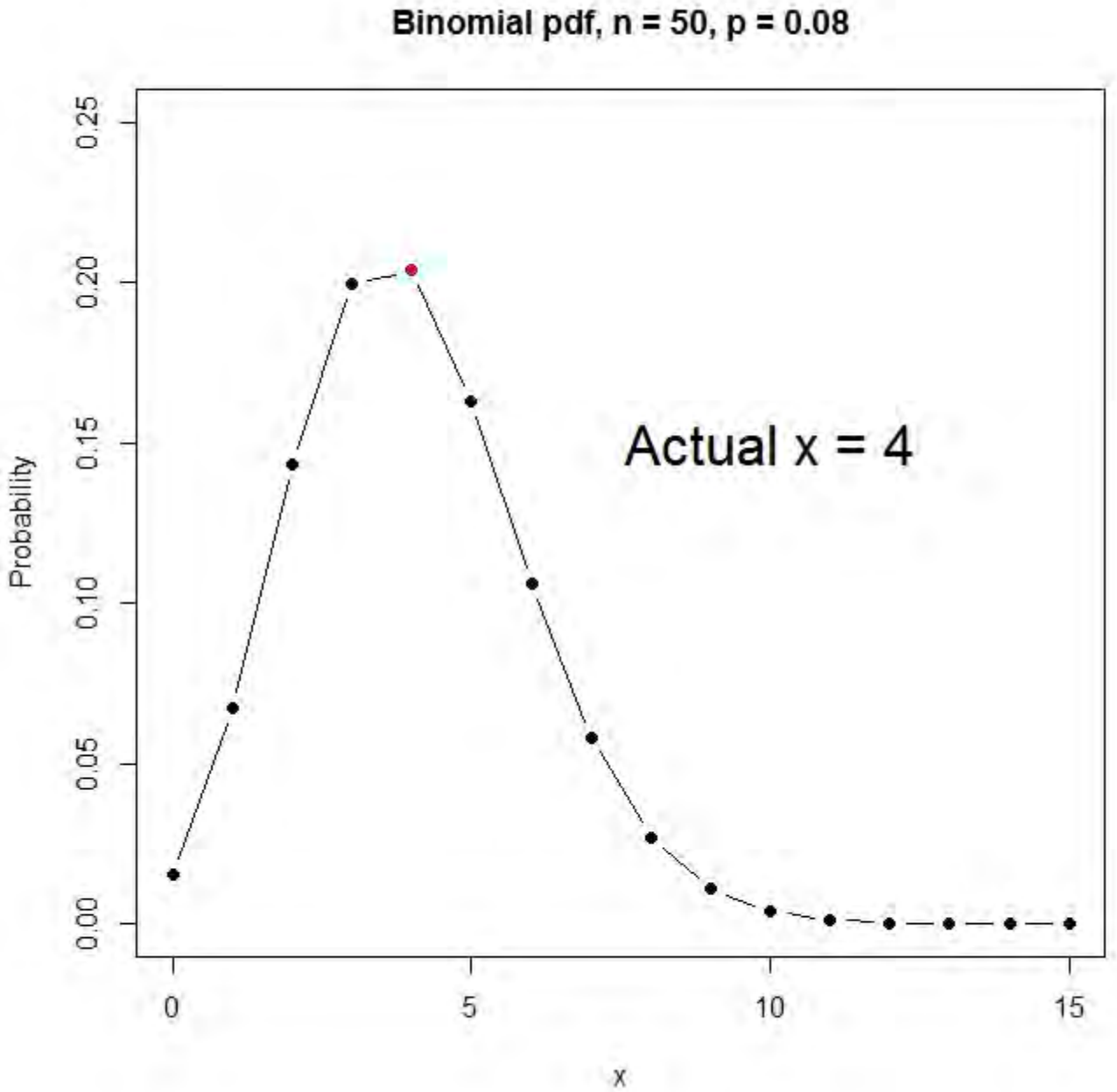
$x(1-p) = p(n-x)$ or $x - xp = pn - px$ or $x = pn$ so

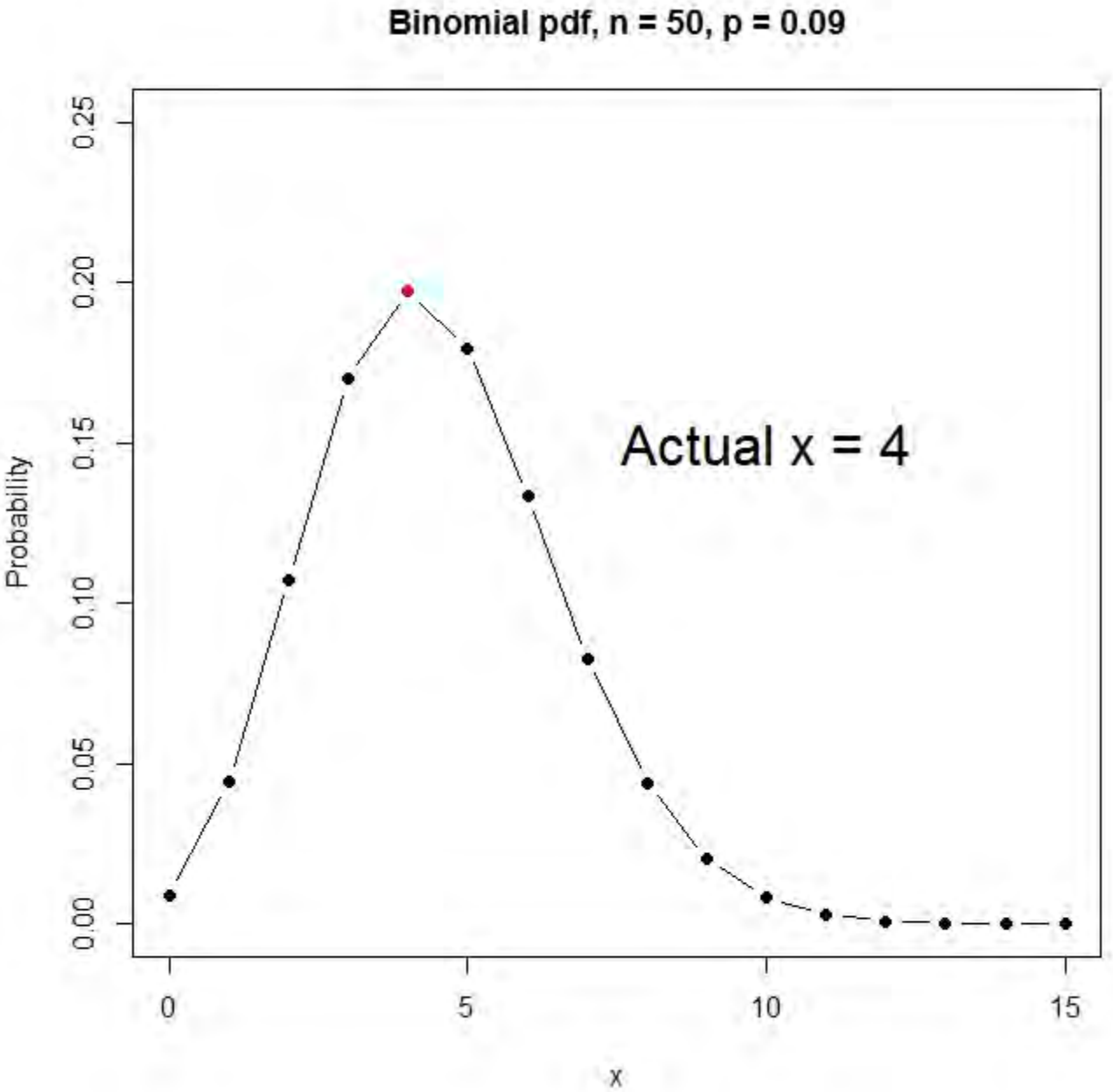
$\hat{p} = x/n$

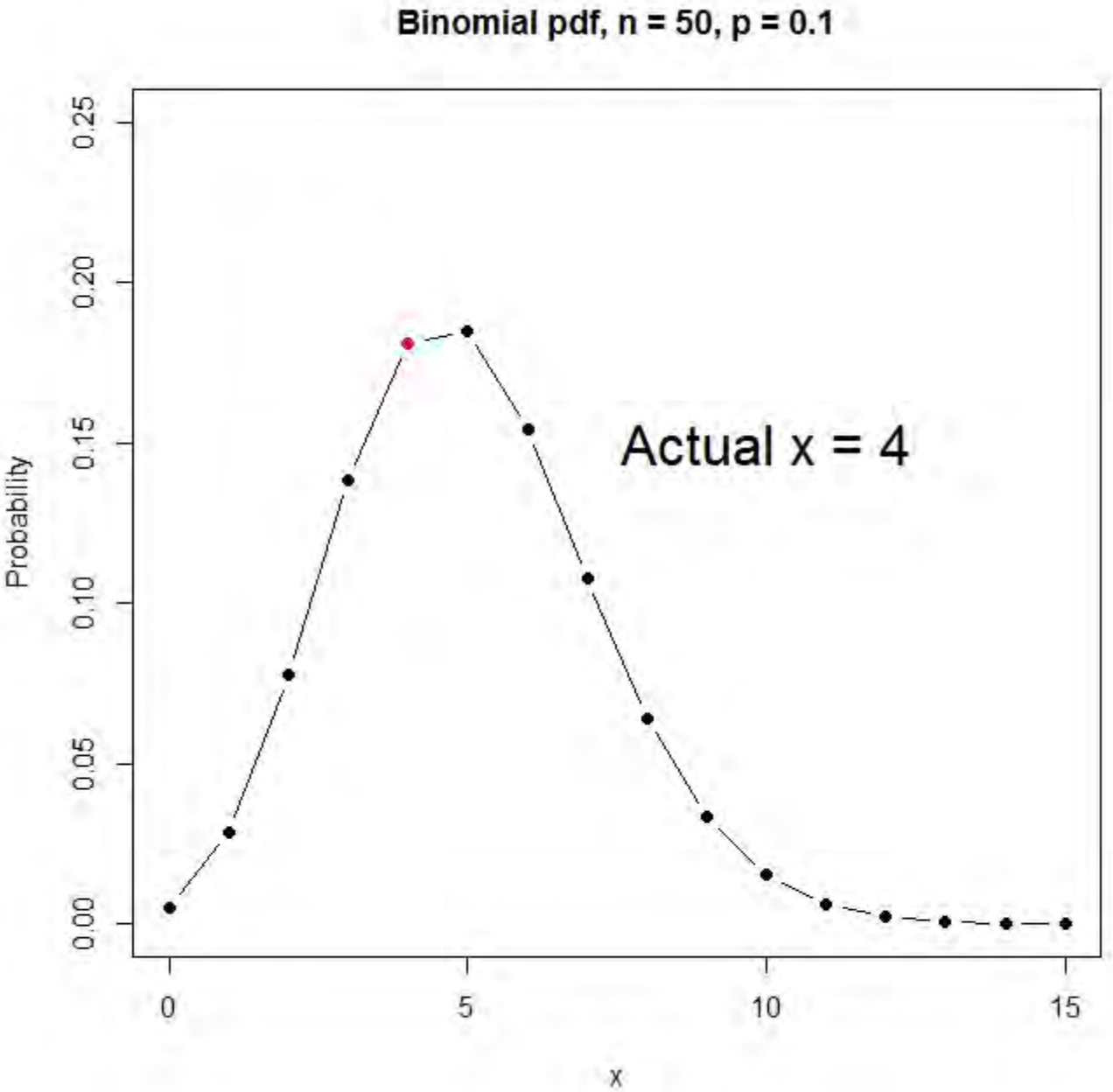
PDF and Likelihood

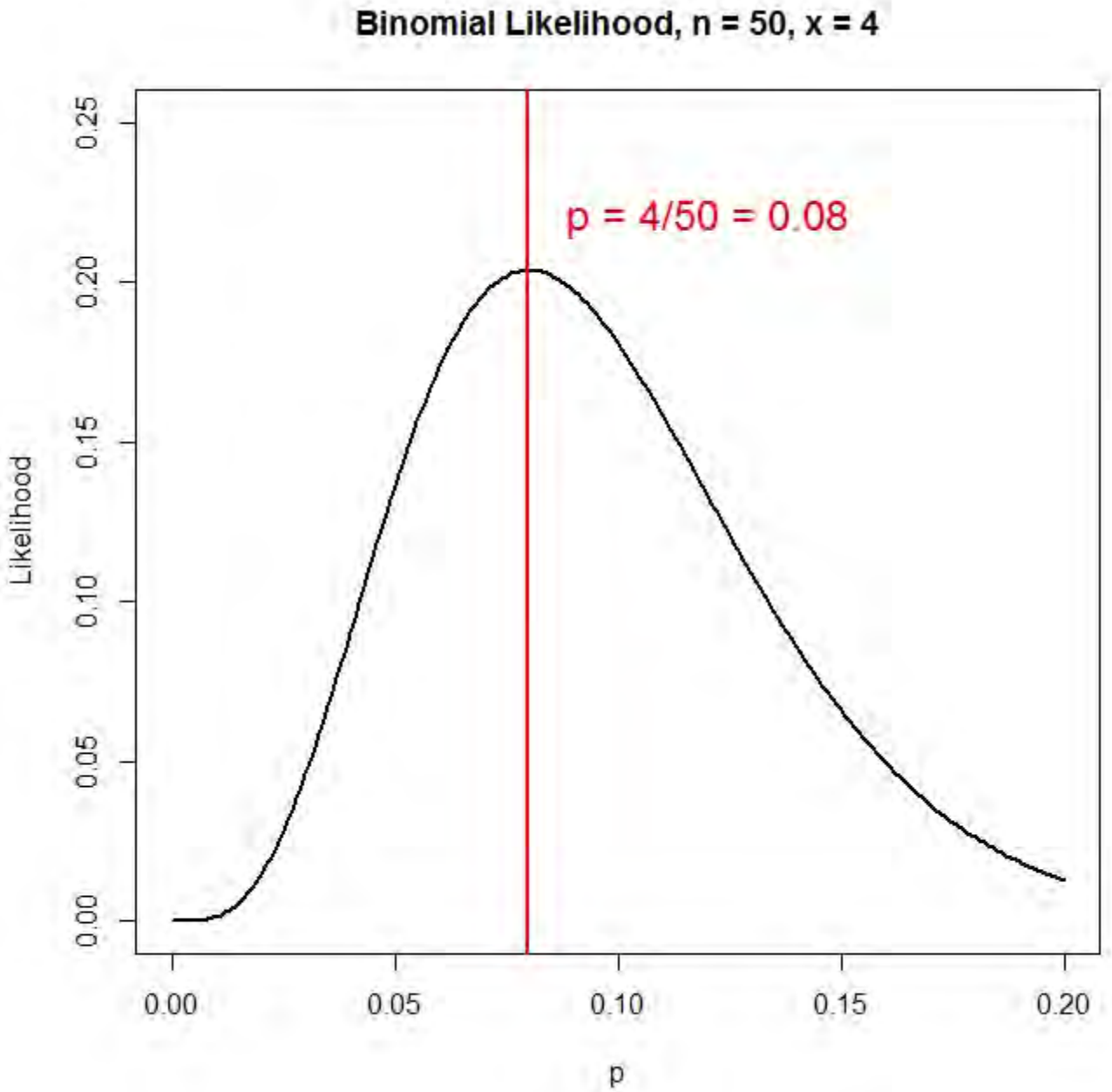
- The next slides show the binomial pdf for $n = 50$ and various values of x, with $x = 4$ highlighted in red.
- The four slides are for $p = 0.07$, 0.08, 0.09, and 0.10.
- To find the MLE for p, we look at the height of the red dot and find the value of p for which it is the highest.
- The MLE is the value of p for which the ex-ante probability of the x value that actually occurred is the highest.


Binomial pdf, n = 50, p = 0.07
Probability
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10
15
x
Actual x = 4


Binomial pdf, n = 50, p = 0.08
Probability
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10
15
x
Actual x = 4


Binomial pdf, n = 50, p = 0.09
Probability
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10
15
x
Actual x = 4


Binomial pdf, n = 50, p = 0.1
Probability
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10
15
x
Actual x = 4


Binomial Likelihood, n = 50, x = 4
p = 4/50 = 0.08
Likelihood
0.00
0.05
0.10
0.15
0.20
0.25
0.00
0.05
0.10
0.15
0.20
p

Two groups, exposed (E) and control (C)

$$\hat{p}_E = x_E / n_E$$

$$\hat{p}_C = x_C / n_C$$

$$w_E = \ln[\hat{p}_E / (1 - \hat{p}_E)]$$

$$w_C = \ln[\hat{p}_C / (1 - \hat{p}_C)]$$

$z = 1$ for exposed $z = 0$ for control

$$\hat{\beta}_0 + \hat{\beta}_1 z$$

$$\hat{\beta}_0 = w_C$$

$$\hat{\beta}_1 = w_E - w_C$$

Most MLE estimates have to be done interatively.

Observations $1 \le i \le n$

Each observation has a covariate z_i and a response $x_i = 0$ or 1.

The linear predictor $\eta_i = \beta_0 + \beta_1 z_i$ with $p_i = [1 + \exp(-\eta_i)]^{-1}$

The likelihood for the ith observation is

$p_i^{x_i}(1-p_i)^{1-x_i}$ which is either p_i or $1-p_i$

according to whether x_i is respectively 1 or 0.

The likelihood is then

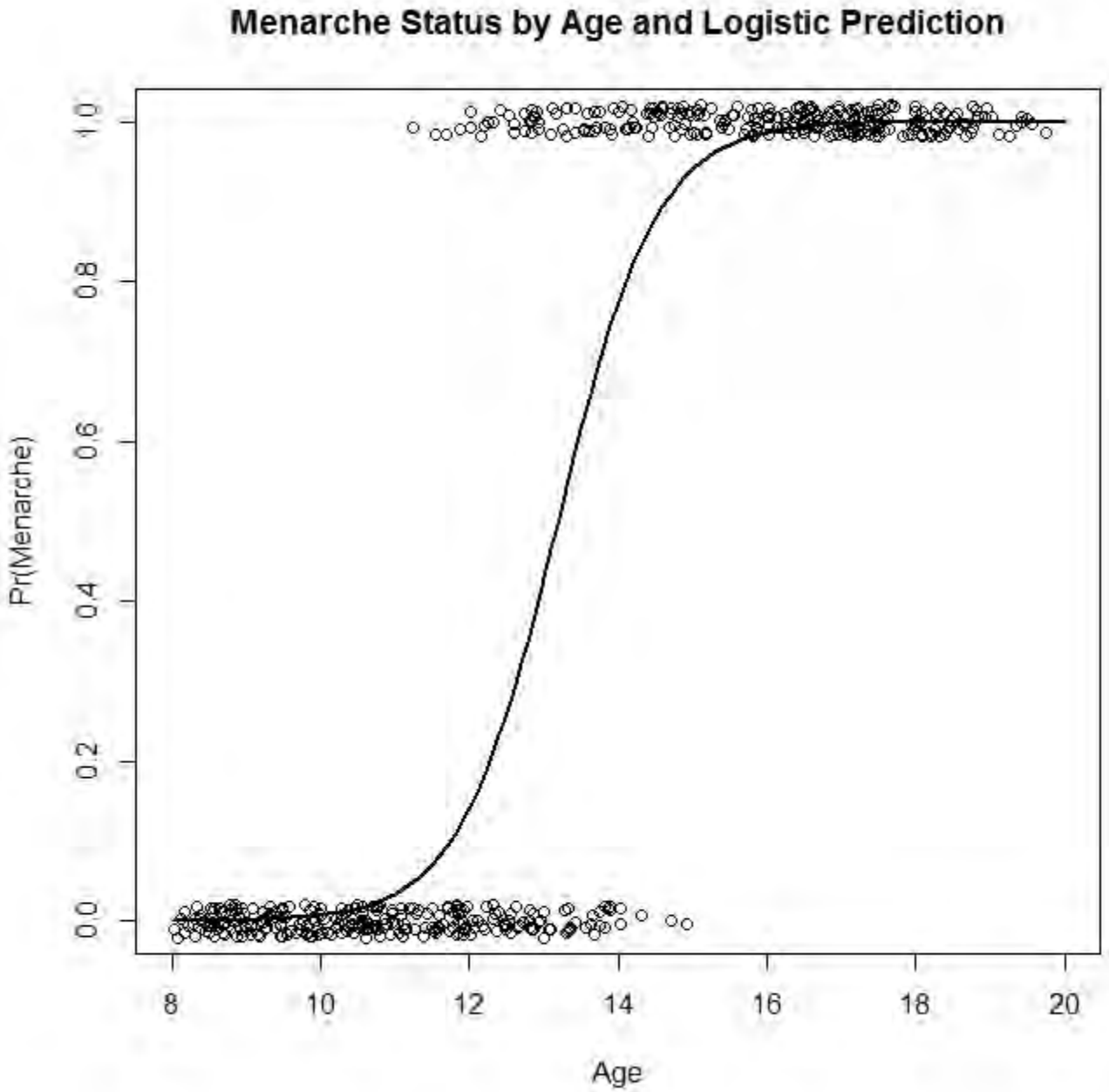
$\prod_{x_i=0}[1-p_i]\prod_{x_i=1}[p_i]$ and the log likelihood is

$$\sum_{x_i=0}\ln(1-p_i)+\sum_{x_i=1}\ln(p_i)$$

and this is maximized numerically with respect

to β_0 and β_1 when they are chosen so that as much as possible

p_i is large when $x_i = 1$ and small when $x_i = 0$


Menarche Status by Age and Logistic Prediction
Pr(Menarche)
0.0
0.2
0.4
0.6
0.8
1.0
8
10
12
14
16
18
20
Age

Suppose that $y_1, y_2, \ldots, y_n \sim \text{Bin}(m, p_i)$

The best fit possible (though maybe not useful) is to set

$\hat{p}_i = y_i / m$ in which case the likelihood is

$$\prod \binom{m}{y_i} (y_i / m)^{y_i} (1 - y_i / m)^{m - y_i}$$

This is the highest possible value for the likelihood and uses n parameters.

If we have a statistical model using fewer parameters that predicts $\hat{p}_i$

and if the likelihood is maximized by choosing the coefficients, and if

we let $\hat{\mu}_i = m\hat{p}_i$ then that likelihood is

$$\prod \binom{m}{y_i} (\hat{\mu}_i / m)^{y_i} (1 - \hat{\mu}_i / m)^{m - y_i}$$

The *deviance* is twice the difference between the best possible log likelihood and the log likelihood under the model.

$$\prod \binom{m}{y_i} (y_i / m)^{y_i} (1 - y_i / m)^{m - y_i}$$

$$\prod \binom{m}{y_i} (\hat{\mu}_i / m)^{y_i} (1 - \hat{\mu}_i / m)^{m - y_i}$$

$$D = 2 \sum \left[y_i \ln(y_i / \hat{\mu}_i) + (m - y_i) \ln((m - y_i) / (m - \hat{\mu}_i)) \right]$$

This is larger if y_i is farther from $\hat{\mu}_i$.

If m varies from observation to observation then just put m_i for m

The *null* model is where all the $\hat{p}_i$ are the same and would be estimated by

$$\hat{p} = \frac{\sum y_i}{mn} \text{ and } \hat{\mu} = m\hat{p}$$

$$ll(\text{null model}) \leq ll(\text{model}) \leq ll(\text{max model})$$

Deviance

- The deviance under the normal distribution is just the residual sum of squares.
- Changes in normal deviance is usually assessed by the F-test in an ANOVA table.
- For logistic regression, differences in deviance are assessed using the chi-squared distribution with degrees of freedom equal to the number of parameters omitted between the larger and smaller model.
- The latter test is approximate (asymptotic).
- Tests of coefficients are based on the Wald test in which we have an estimate and an estimated variance. This too is approximate and not identical to the likelihood ratio (deviance) test except in linear regression.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.37766    0.38018  -6.254    4e-10 ***
smokingYes  -0.06777    0.27812  -0.244   0.8075
obesityYes   0.69531    0.28509   2.439   0.0147 *
snoringYes   0.87194    0.39757   2.193   0.0283 *
---

    Null deviance: 14.1259  on 7  degrees of freedom
Residual deviance:  1.6184  on 4  degrees of freedom
AIC: 34.537

> deviance(glm(hyp.tbl ~ smoking+obesity+snoring,family=binomial))
[1] 1.618403
> deviance(glm(hyp.tbl ~ smoking*obesity*snoring,family=binomial))
[1] 4.525669e-10
> deviance(glm(hyp.tbl ~ 1,family=binomial))
[1] 14.1259
> logLik(glm(hyp.tbl ~ smoking+obesity+snoring,family=binomial))
'log Lik.' -13.26858 (df=4)
> logLik(glm(hyp.tbl ~ smoking*obesity*snoring,family=binomial))
'log Lik.' -12.45938 (df=8)
> 2*(13.26858-12.45938)
[1] 1.6184
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.37766    0.38018  -6.254    4e-10 ***
smokingYes  -0.06777    0.27812  -0.244   0.8075
obesityYes   0.69531    0.28509   2.439   0.0147 *
snoringYes   0.87194    0.39757   2.193   0.0283 *
---

    Null deviance: 14.1259  on 7  degrees of freedom
Residual deviance:  1.6184  on 4  degrees of freedom
AIC: 34.537

> extractAIC(glm(hyp.tbl ~ smoking+obesity+snoring,family=binomial))
[1]  4.00000 34.53717
> extractAIC(glm(hyp.tbl ~ obesity+snoring,family=binomial))
[1]  3.00000 32.59689
> extractAIC(glm(hyp.tbl ~ smoking+snoring,family=binomial))
[1]  3.00000 38.19373
```

```
> drop1(hyp.glm,test="Chisq")
Single term deletions

Model:
n.hyp.n.tot ~ smoking + obesity + snoring
        Df Deviance    AIC    LRT Pr(>Chi)
<none>       1.6184 34.537
smoking  1   1.6781 32.597 0.0597  0.80694
obesity  1   7.2750 38.194 5.6566  0.01739 *
snoring  1   7.2963 38.215 5.6779  0.01718 *
---
```

Deviance = (llhd – llhd(perfect fit))
AIC = - 2*log L + k * df

The AIC has a penalty for more parameters because otherwise, the llhd would Always increase when variables are added. We like small AIC, small deviance, and large llhd.

```
> coef(summary(hyp.glm))
               Estimate Std. Error    z value     Pr(>|z|)
(Intercept) -2.37766146  0.3801845 -6.2539671 4.001553e-10
smokingYes  -0.06777489  0.2781242 -0.2436857 8.074742e-01
obesityYes   0.69530960  0.2850851  2.4389544 1.472983e-02
snoringYes   0.87193932  0.3975736  2.1931517 2.829645e-02
> vcov(hyp.glm)
            (Intercept)    smokingYes    obesityYes   snoringYes
(Intercept)  0.14454027 -1.607354e-02 -1.474522e-02 -0.135505811
smokingYes  -0.01607354  7.735305e-02 -8.029255e-06 -0.007415799
obesityYes  -0.01474522 -8.029255e-06  8.127352e-02 -0.008143230
snoringYes  -0.13550581 -7.415799e-03 -8.143230e-03  0.158064803

> round(vcov(hyp.glm),4)
            (Intercept) smokingYes obesityYes snoringYes
(Intercept)      0.1445    -0.0161    -0.0147    -0.1355
smokingYes      -0.0161     0.0774     0.0000    -0.0074
obesityYes      -0.0147     0.0000     0.0813    -0.0081
snoringYes      -0.1355    -0.0074    -0.0081     0.1581

>sqrt(diag(vcov(hyp.glm)))
(Intercept)  smokingYes  obesityYes  snoringYes
  0.3801845   0.2781242   0.2850851   0.3975736
```

Wald Tests of Coefficients

- The variance-covariance matrix of the coefficients in glm is based on approximate (asymptotic) theory.
- The estimates will be better for larger sample size.
- This can be used to get a CI for a coefficient or for a difference of coefficients.
- The test and p-values should be similar to the results from the likelihood ratio test but will not be identical.
- It is worth looking at both.
- **But the likelihood ratio test may be better.**

Homework 2a: Due 4/8/21

- In 1973, a large cotton textile company in North Carolina made a study to investigate the prevalence of byssinosis, a form of pneumoconiosis to which workers exposed to cotton dust are subject.
- We will investigate relationships between disease and sex, race, length of employment, smoking, and dustiness of workplace.
- There are 5,419 workers in the data set.

Data

Variable	Description
Type of work place	1 (most dusty), 2 (less dusty), 3 (least dusty)
Employment, years	< 10, 10–19, 20–
Smoking	Smoker or not in last 5 years
Sex	Male, Female
Race	White, other
Byssinosis	Yes/No

Assignment

- Read the data into R, SAS, or another statistical package.
- How many different groups (combinations) of exposure and control factors are there?
- Fit a logistic regression model using all these factors.
- Which ones appear statistically significant? Use both the Wald and Likelihood ratio test and explain which is which.
- Compute the estimated odds ratios for each factor and for comparisons within Employment and Workspace, and also compute confidence intervals if you can.
- Which factors appear most important?
- Interactions, if any, will be left to later.