Roberta De Vito

The coefficient: a summary

$$\text{Percent Change in the Odds} = (e^{\beta_1} - 1) \times 100$$

```
> summary(fit.1)

Call:
glm(formula = vote ~ income, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2756  -1.0034  -0.8796   1.2194   1.6550

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.40213    0.18946  -7.401 1.35e-13 ***
income       0.32599    0.05688   5.731 9.97e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1591.2  on 1178  degrees of freedom
Residual deviance: 1556.9  on 1177  degrees of freedom
  (368 observations deleted due to missingness)
AIC: 1560.9

Number of Fisher Scoring iterations: 4
```

The income coefficient

How much increase in the odds do we have with a coefficient of 0.326?

1. 38.5%
2. 32.6 %
3. 61.5%

Adding some covariates

```
> summary(fit.2)

Call:
glm(formula = vote ~ income + black2, family = binomial(link = "logit"))

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.367  -1.185  -0.352   1.170   2.372

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.60855    0.21181  -2.873  0.00407 **
income       0.20886    0.06551   3.188  0.00143 **
black2      -2.55929    0.43030  -5.948 2.72e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1188.1  on 860  degrees of freedom
Residual deviance: 1101.2  on 858  degrees of freedom
  (323 observations deleted due to missingness)
AIC: 1107.2

Number of Fisher Scoring iterations: 5
```

The income coefficient in model 2

How much increase in the odds do we have with a coefficient of 0.209?

1. 76.7%
2. 23.3 %
3. 20.9%

The black coefficient in model 2

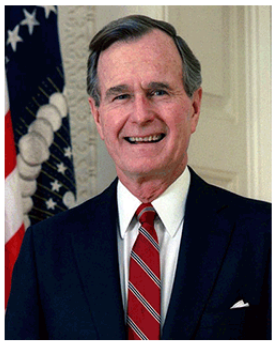
How much decrease in the odds do we have with a coefficient of -2.56?

1. 7 %
2. 92.3 %
3. 56 %

Groups voote in 1992

Bill Clinton/Al Gore

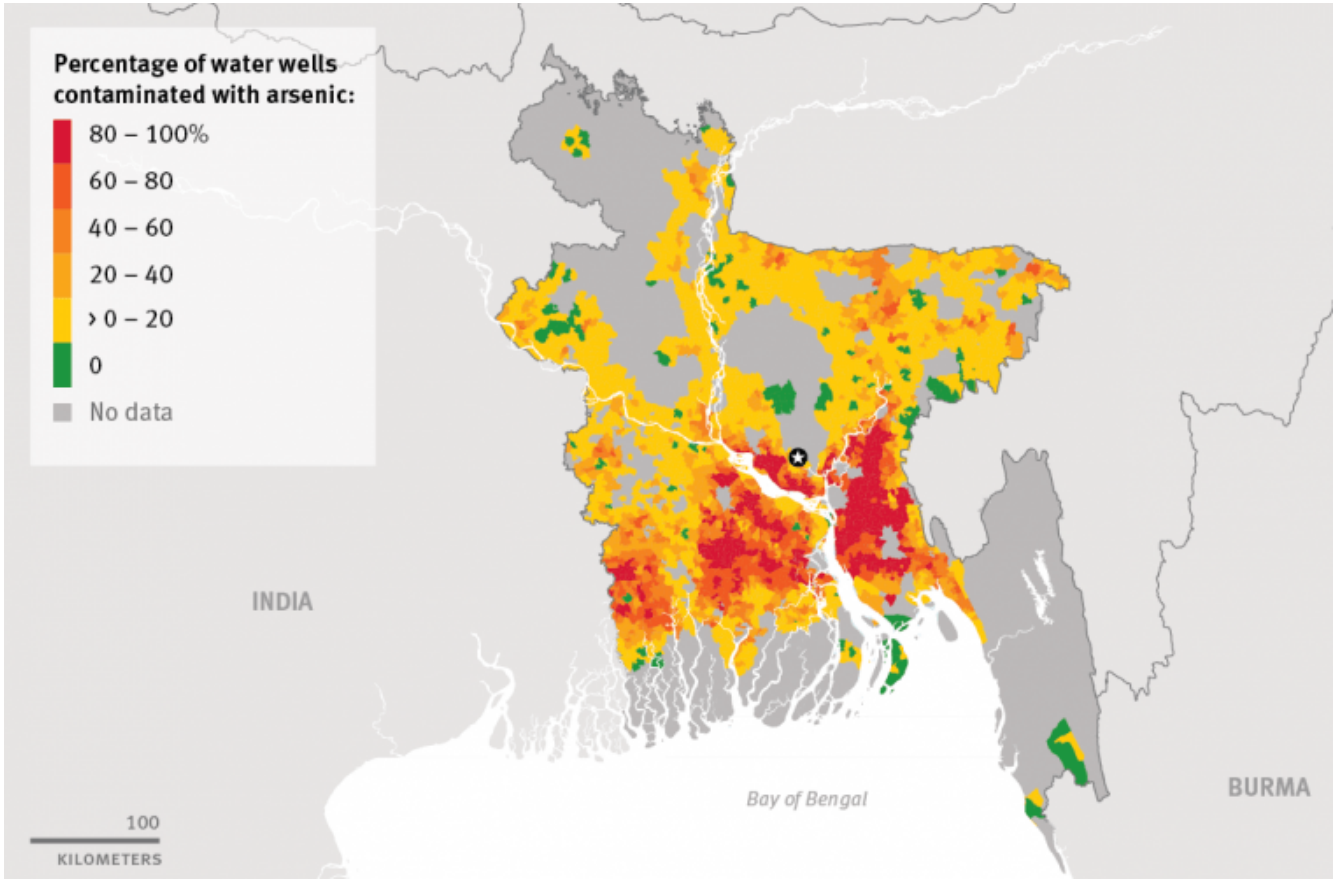
UNITED STATES
ELECTIONS



George H.W. Bush/Dan Quayle

	1992	GROUP	CLINTON	BUSH	PEROT
	All Voters	Pct.	43%	37%	19%
SEX	Men	47	41	38	21
	Women	53	45	38	17
RACE	White	87	39	41	21
	African-American	8	83	10	7
	Hispanic	2	61	25	14
	Asian	1	31	55	15

Arsenic level in Bangladesh



Mathematical model

- the outcome

$$y_i = \begin{cases} 1 & \text{if household i switched to a new well} \\ 0 & \text{if household i continued using its own well} \end{cases}$$

- The distance (in meters) to the closest known safe well
- The arsenic level of respondent's well
- Whether any members of the household are active in community organizations
- The education level of the head of household.

Logistic regression with just one predictor

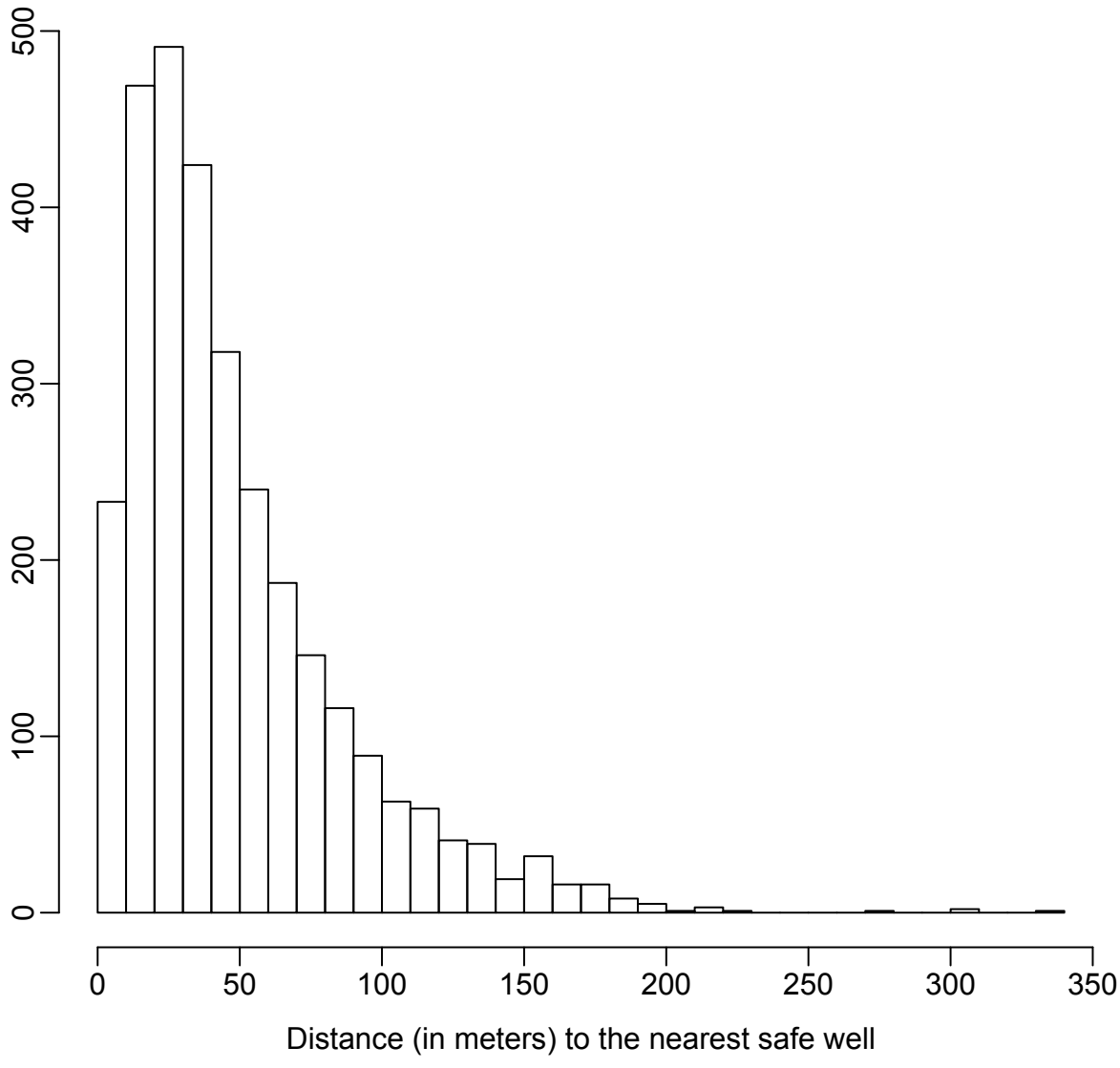
```
> summary(fit.1)

Call:
glm(formula = switch ~ dist, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4406  -1.3058   0.9669   1.0308   1.6603

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.6059594  0.0603102  10.047  < 2e-16 ***
dist        -0.0062188  0.0009743  -6.383 1.74e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4118.1  on 3019  degrees of freedom
Residual deviance: 4076.2  on 3018  degrees of freedom
AIC: 4080.2

Number of Fisher Scoring iterations: 4
```

The distance coefficient in model 1

How much decrease in the odds do we have with a coefficient of -0.0062?

1. 0.62%
2. 6.2 %
3. 62 %

Histogram of distance to the nearest safe well



Histogram of distance to the nearest safe well

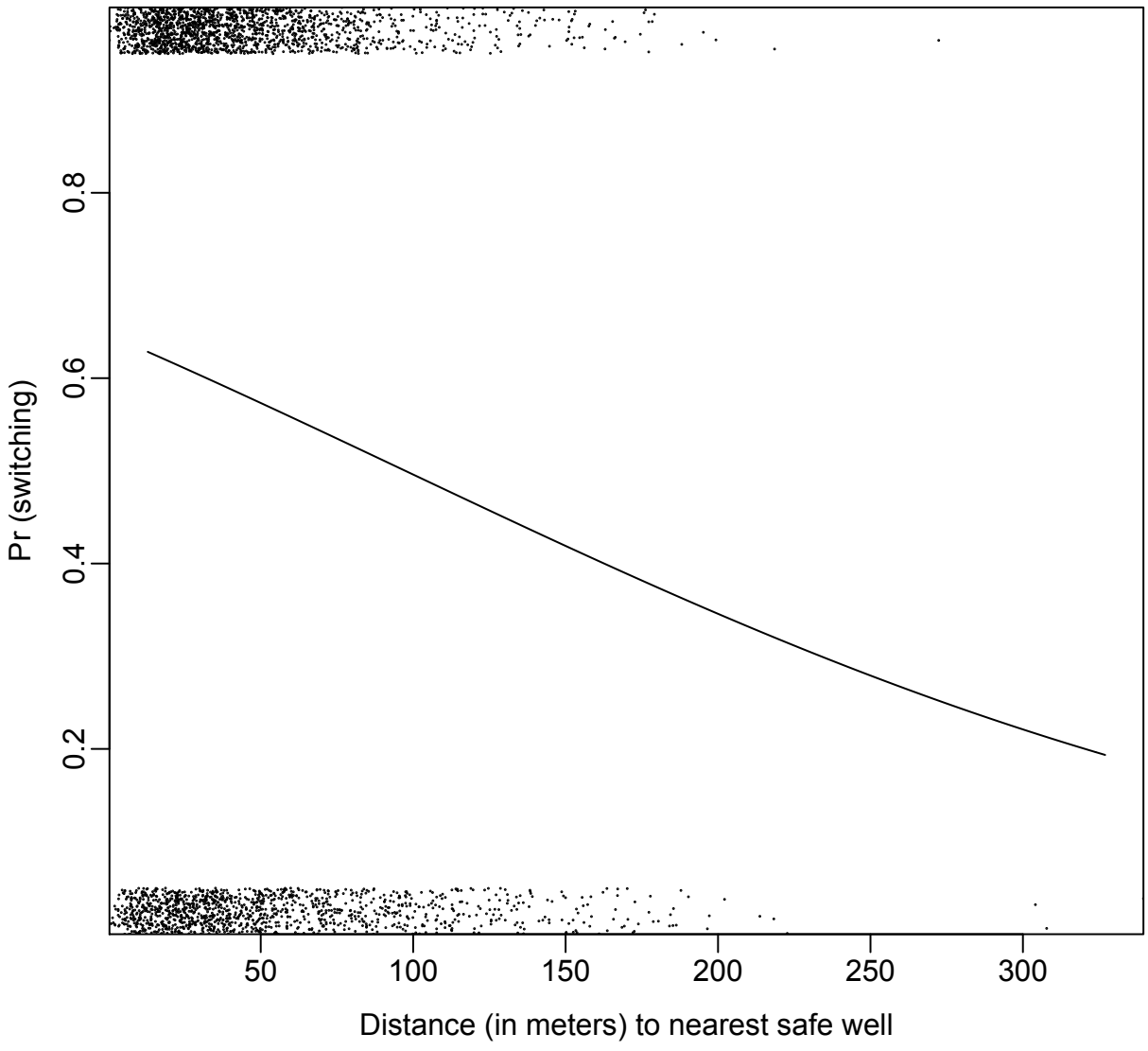
```
> dist100 <- dist/100
> fit.2 <- glm (switch ~ dist100, family=binomial(link="logit"))
> summary(fit.2)

Call:
glm(formula = switch ~ dist100, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4406  -1.3058   0.9669   1.0308   1.6603

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.60596    0.06031  10.047  < 2e-16 ***
dist100     -0.62188    0.09743  -6.383 1.74e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4118.1  on 3019  degrees of freedom
Residual deviance: 4076.2  on 3018  degrees of freedom
AIC: 4080.2

Number of Fisher Scoring iterations: 4
```

The distance coefficient in model 2

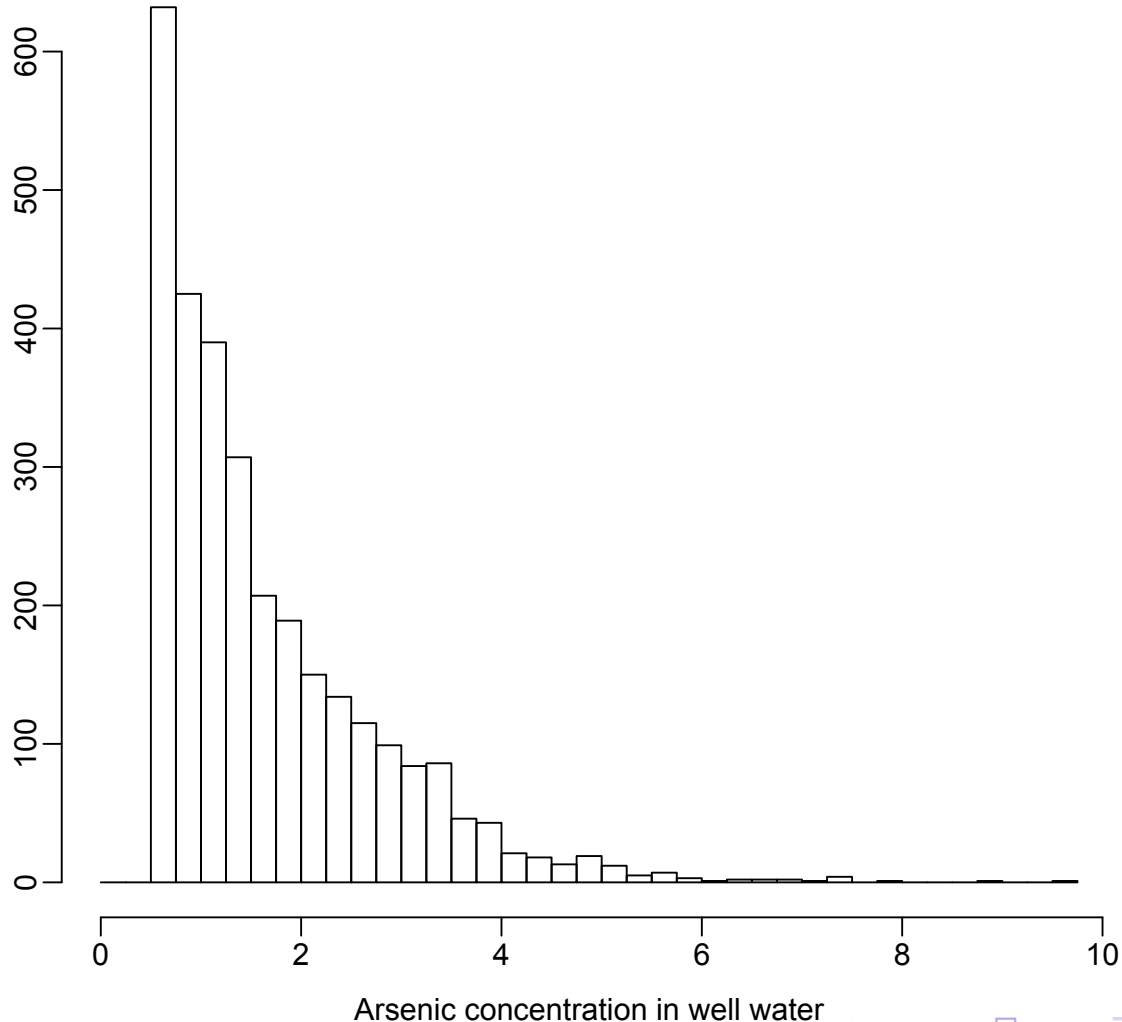
How much decrease in the odds do we have with a coefficient of -0.62?

1. 0.62 %
2. 46 %
3. 6.2 %

Graphical expression of the fitted logistic regression



Histogram of arsenic levels in unsafe wells (those exceeding 0.5)



Logistic regression model with arsenic

```
> summary(fit.3)

Call:
glm(formula = switch ~ dist100 + arsenic, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6351  -1.2139   0.7786   1.0702   1.7085

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.002749   0.079448   0.035    0.972
dist100     -0.896644   0.104347  -8.593   <2e-16 ***
arsenic      0.460775   0.041385  11.134   <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: 4118.1  on 3019  degrees of freedom
Residual deviance: 3930.7  on 3017  degrees of freedom
AIC: 3936.7

Number of Fisher Scoring iterations: 4
```

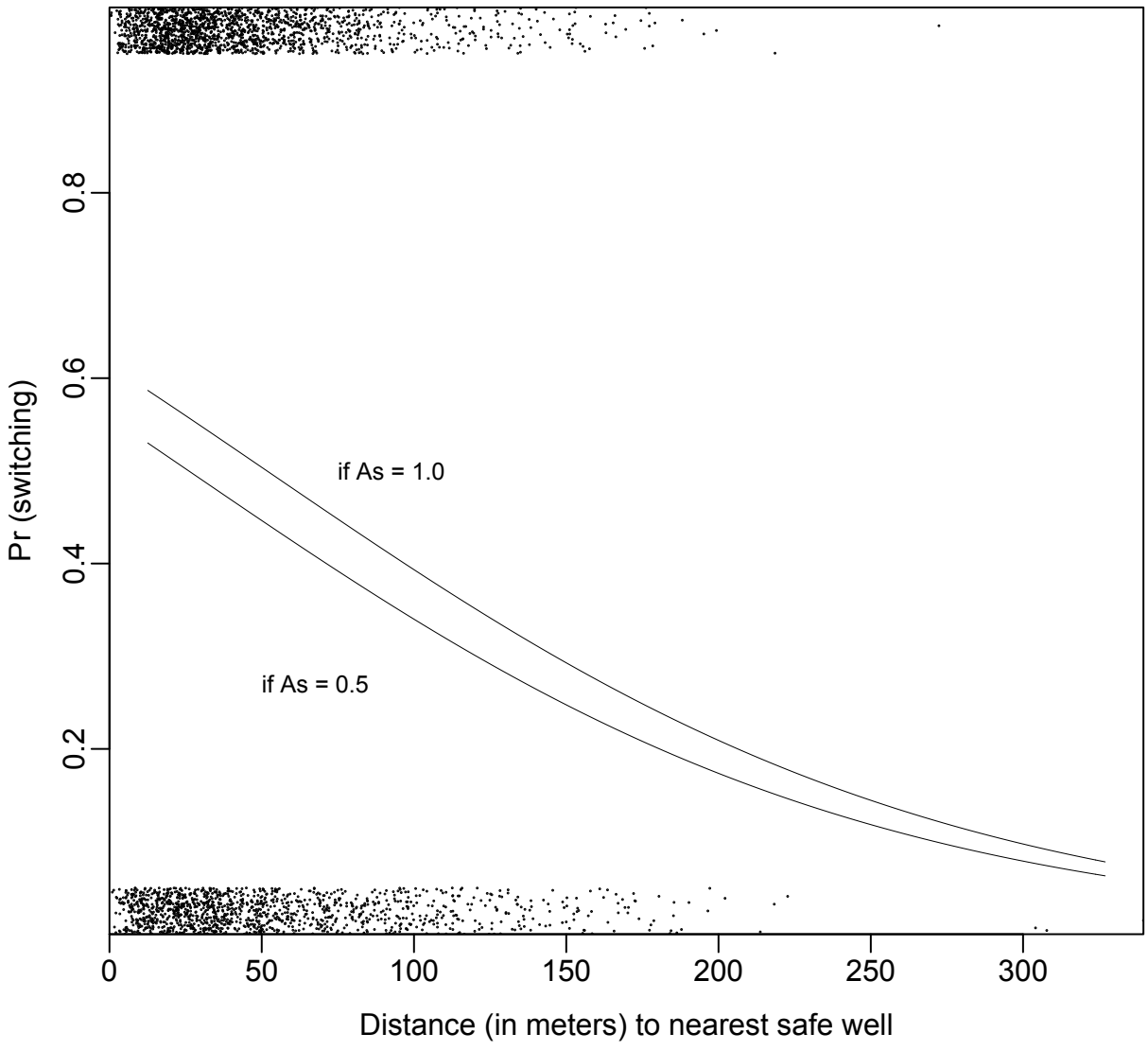
The arsenic coefficient in model 3

How much increase in the odds we have with a coefficient of 0.46?

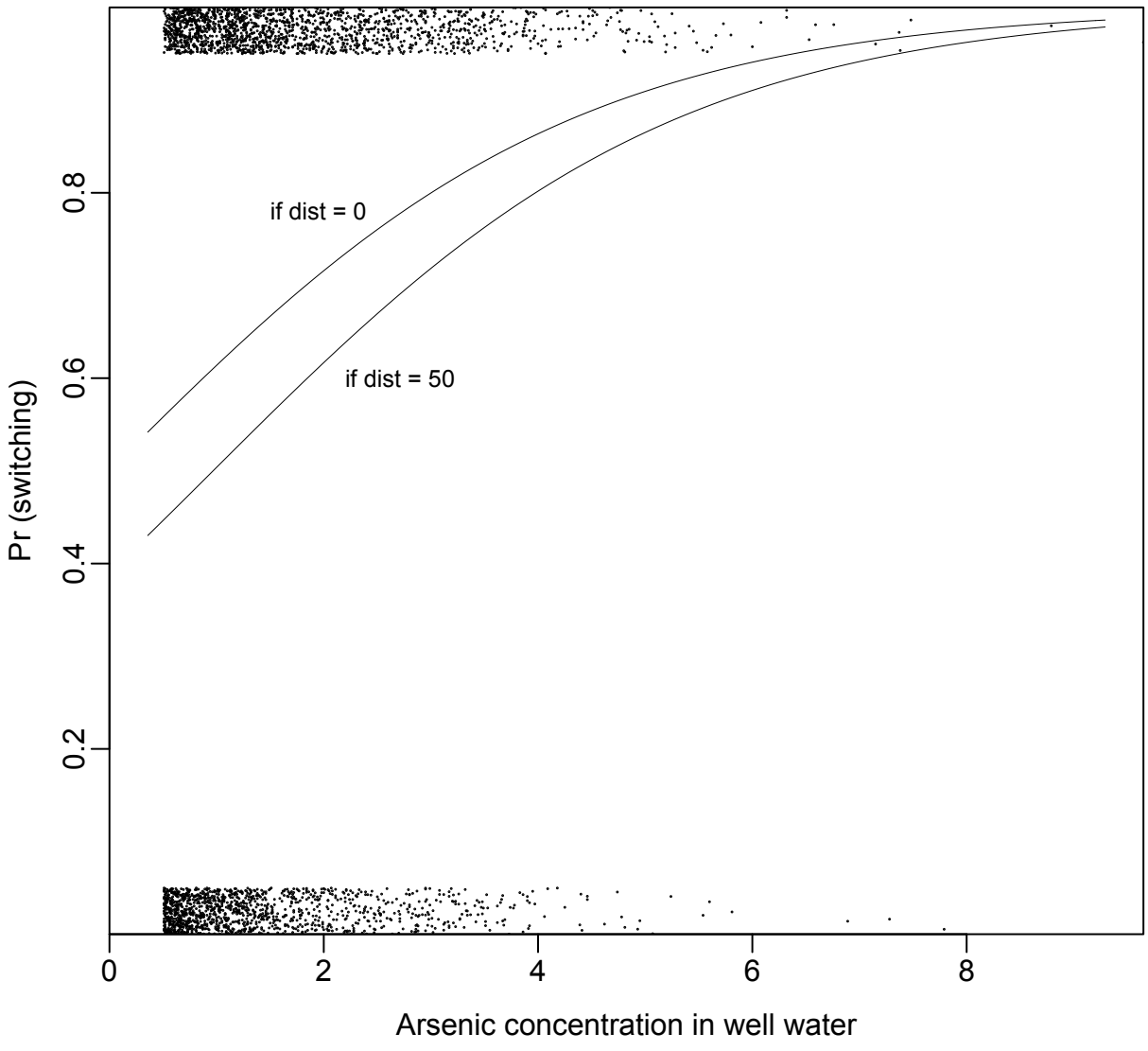
1. 0.46 %
2. 58.4 %
3. 46 %

Another question: go to prismia!

Fitted logistic regression of probability of switching



Fitted logistic regression of probability of switching II



Logistic regression with interactions

```
> summary(fit.4)

Call:
glm(formula = switch ~ dist100 + arsenic + dist100:arsenic, family = binomial(link =
"logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7823  -1.2004   0.7696   1.0816   1.8476

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.14787    0.11754  -1.258  0.20838
dist100        -0.57722    0.20918  -2.759  0.00579 **
arsenic         0.55598    0.06932   8.021 1.05e-15 ***
dist100:arsenic -0.17891    0.10233  -1.748  0.08040 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4118.1  on 3019  degrees of freedom
Residual deviance: 3927.6  on 3016  degrees of freedom
AIC: 3935.6

Number of Fisher Scoring iterations: 4
```

Each regression coefficient in turn

- Constant term: $exp^{-0.148} - 1 = -0.14$
- Coefficient for distance: $exp^{-0.577} - 1 = -0.439$
- Coefficient for arsenic: $exp^{0.556} - 1 = 0.74$
- Coefficient for the interaction term: $exp^{-0.18} - 1 = -0.16$

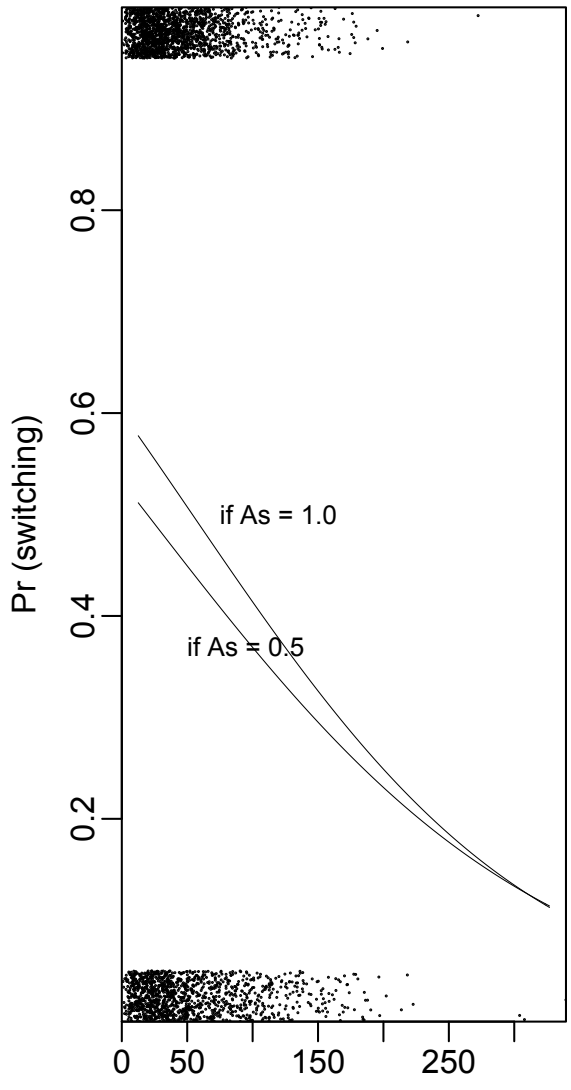
Logistic regression: centering the variables

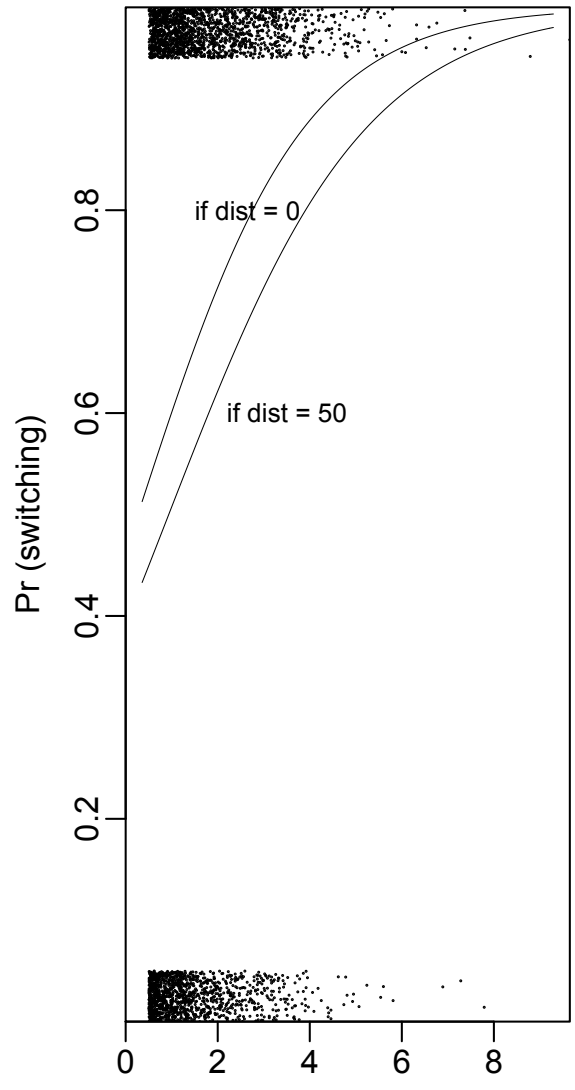
```
> c.dist100 <- dist100 - mean (dist100)
> c.arsenic <- arsenic - mean (arsenic)
> fit.5 <- glm (switch ~ c.dist100 + c.arsenic + c.dist100:c.arsenic,
+   family=binomial(link="logit"))
> summary(fit.5)

Call:
glm(formula = switch ~ c.dist100 + c.arsenic + c.dist100:c.arsenic,
    family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7823  -1.2004   0.7696   1.0816   1.8476

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.35109    0.03985   8.810   <2e-16 ***
c.dist100          -0.87365    0.10480  -8.337   <2e-16 ***
c.arsenic           0.46951    0.04207  11.159   <2e-16 ***
c.dist100:c.arsenic -0.17891    0.10233  -1.748   0.0804 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4118.1  on 3019  degrees of freedom
Residual deviance: 3927.6  on 3016  degrees of freedom
AIC: 3935.6

Number of Fisher Scoring iterations: 4
```

Interpreting the inferences on this new scale:

- the constant term: from 14% to 41%
- Coefficient for distance: from -44% to -58%
- Coefficient for arsenic: from 74 % to 59%
- Coefficient for the interaction term: the same

Fitted logistic regression of probability of switching





Logistic regression: adding social predictors

```
> summary(fit.6)

Call:
glm(formula = switch ~ c.dist100 + c.arsenic + c.dist100:c.arsenic +
    assoc + educ4, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7303  -1.1892   0.7444   1.0675   1.6987

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.20252    0.06930   2.922  0.00347 **
c.dist100          -0.87528    0.10507  -8.330  < 2e-16 ***
c.arsenic           0.47531    0.04229  11.238  < 2e-16 ***
assoc              -0.12319    0.07698  -1.600  0.10953
educ4               0.16779    0.03838   4.372 1.23e-05 ***
c.dist100:c.arsenic -0.16123    0.10225  -1.577  0.11482
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4118.1  on 3019  degrees of freedom
Residual deviance: 3905.4  on 3014  degrees of freedom
AIC: 3917.4

Number of Fisher Scoring iterations: 4
```

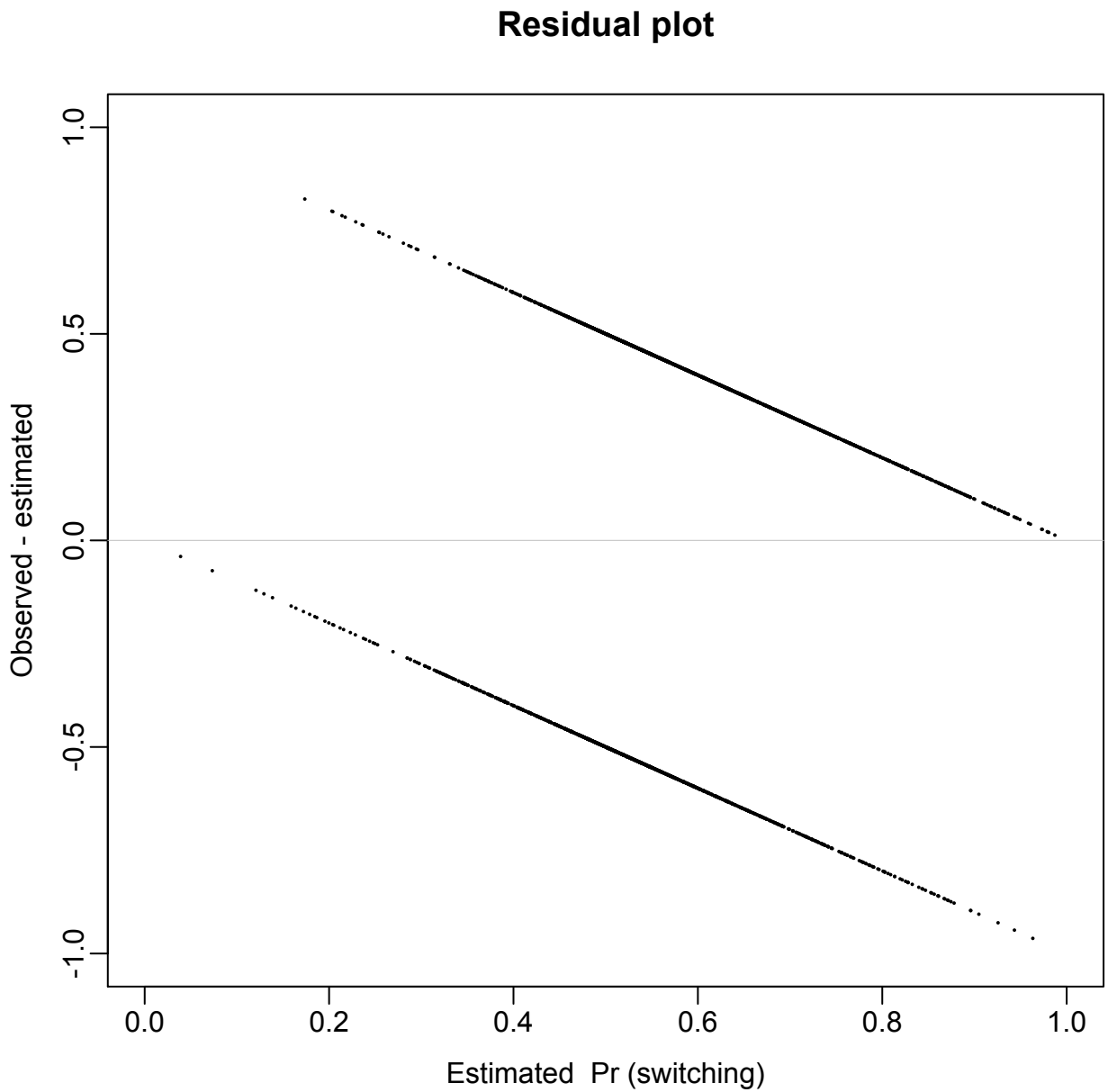
Logistic regression: adding social predictors II

```
> summary (fit.7)

Call:
glm(formula = switch ~ c.dist100 + c.arsenic + c.dist100:c.arsenic +
    educ4, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7149  -1.1886   0.7478   1.0689   1.7223

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.14844    0.06044   2.456   0.0141 *
c.dist100          -0.87462    0.10510  -8.322  < 2e-16 ***
c.arsenic           0.47663    0.04228  11.273  < 2e-16 ***
educ4               0.16922    0.03833   4.415 1.01e-05 ***
c.dist100:c.arsenic -0.16291    0.10235  -1.592   0.1115
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4118.1  on 3019  degrees of freedom
Residual deviance: 3907.9  on 3015  degrees of freedom
AIC: 3917.9

Number of Fisher Scoring iterations: 4
```

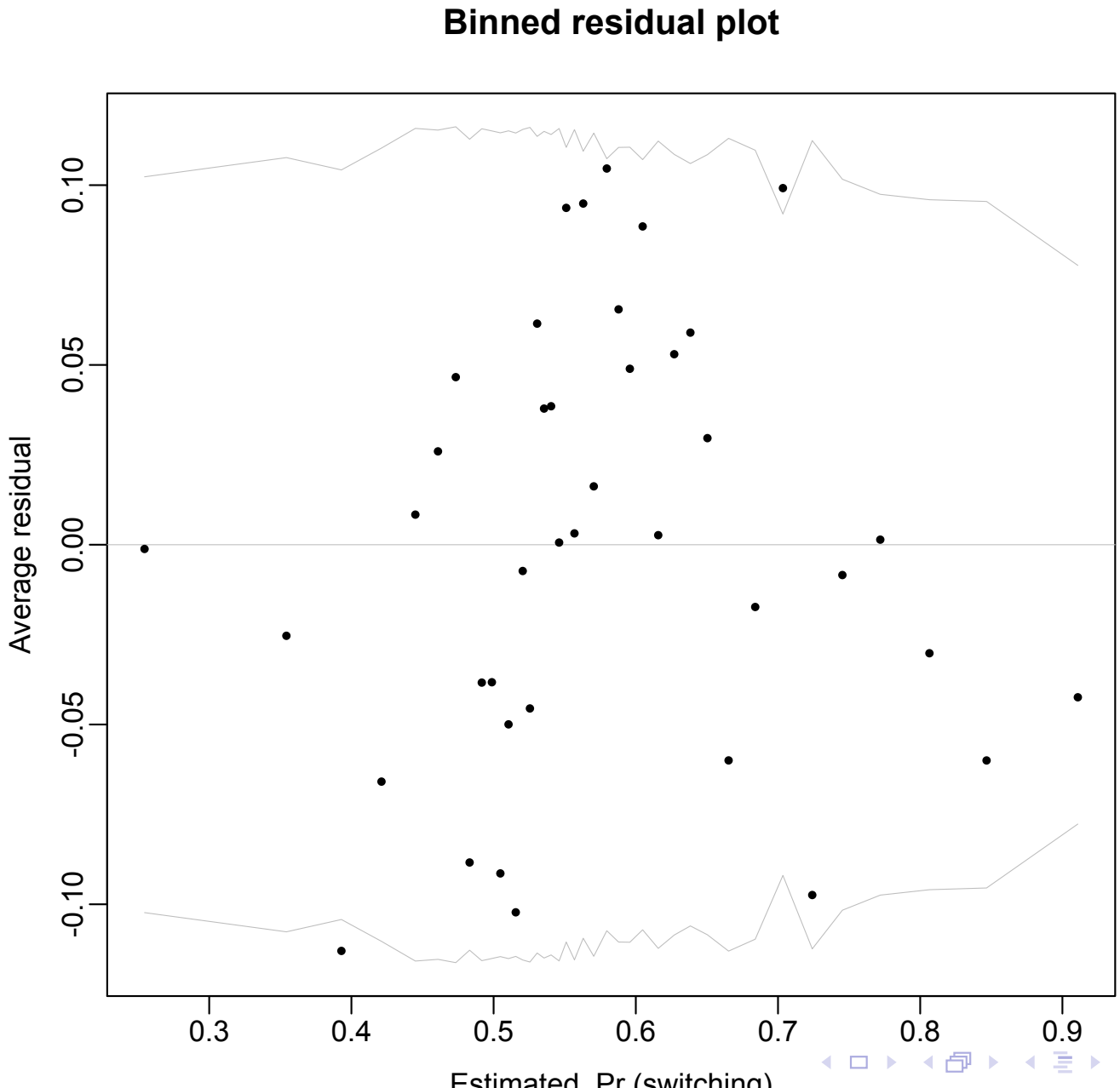
Logistic regression: adding social predictors III

```
> summary (fit.8)

Call:
glm(formula = switch ~ c.dist100 + c.arsenic + c.educ4 + c.dist100:c.arsenic +
    c.dist100:c.educ4 + c.arsenic:c.educ4, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5706  -1.1964   0.7314   1.0724   1.8712

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.35630    0.04028   8.844  < 2e-16 ***
c.dist100          -0.90286    0.10731  -8.414  < 2e-16 ***
c.arsenic           0.49498    0.04305  11.497  < 2e-16 ***
c.educ4             0.18498    0.03919   4.720 2.36e-06 ***
c.dist100:c.arsenic -0.11768    0.10353  -1.137  0.25569
c.dist100:c.educ4    0.32269    0.10662   3.026  0.00247 **
c.arsenic:c.educ4    0.07223    0.04387   1.647  0.09965 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4118.1  on 3019  degrees of freedom
Residual deviance: 3891.7  on 3013  degrees of freedom
AIC: 3905.7

Number of Fisher Scoring iterations: 4
```

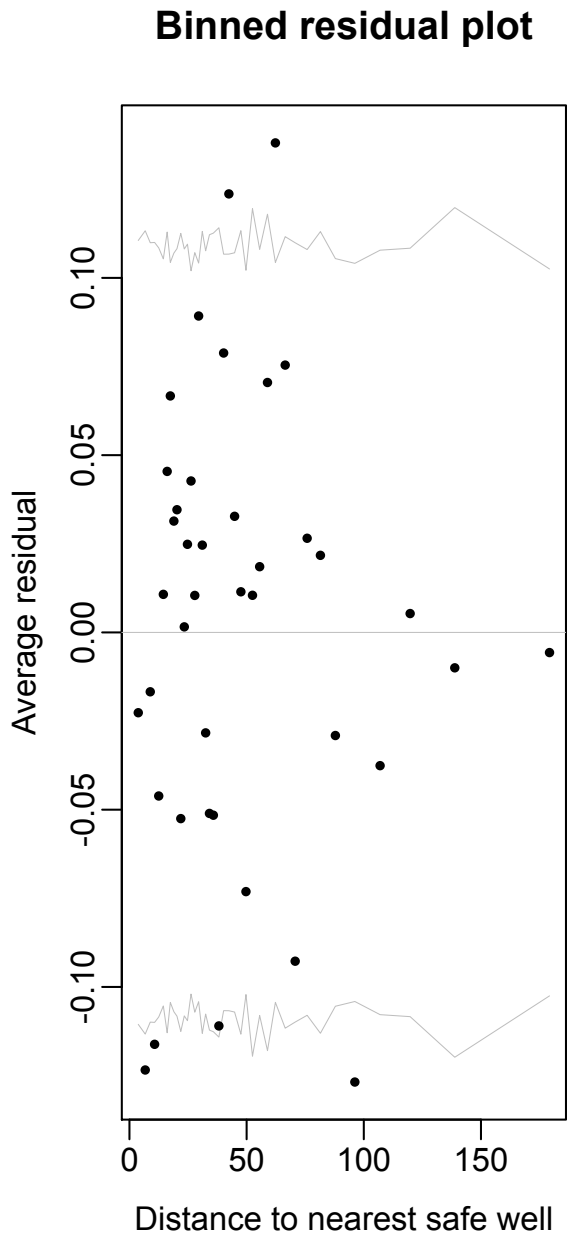
Evaluating, Checking: Residuals

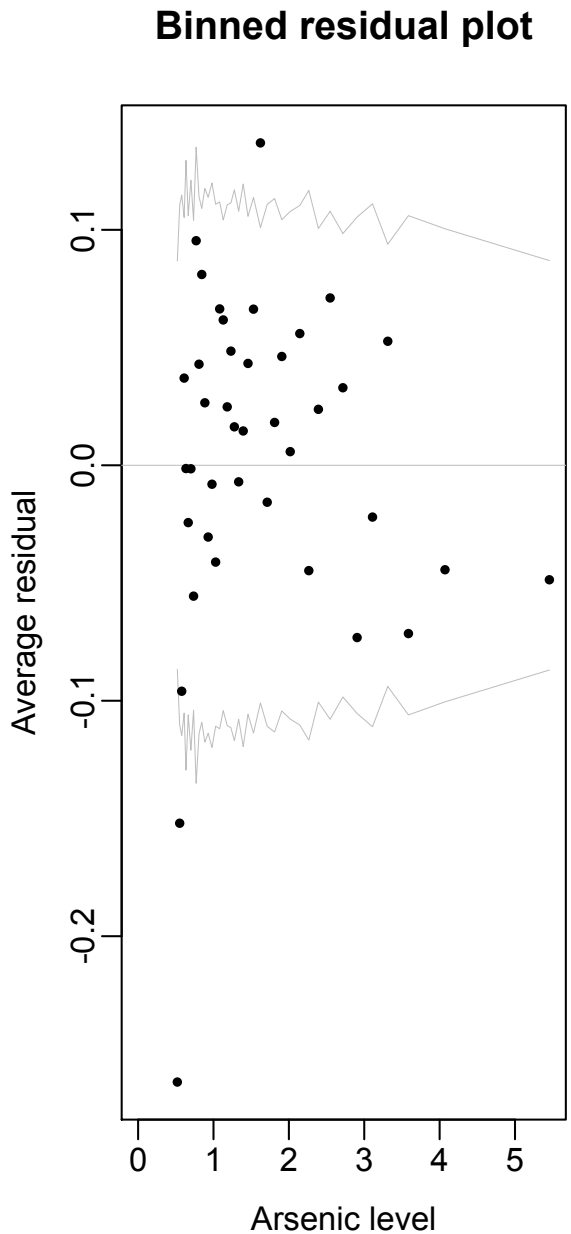


Evaluating, Checking: Binned Residuals



Evaluating, Checking: Binned Residuals

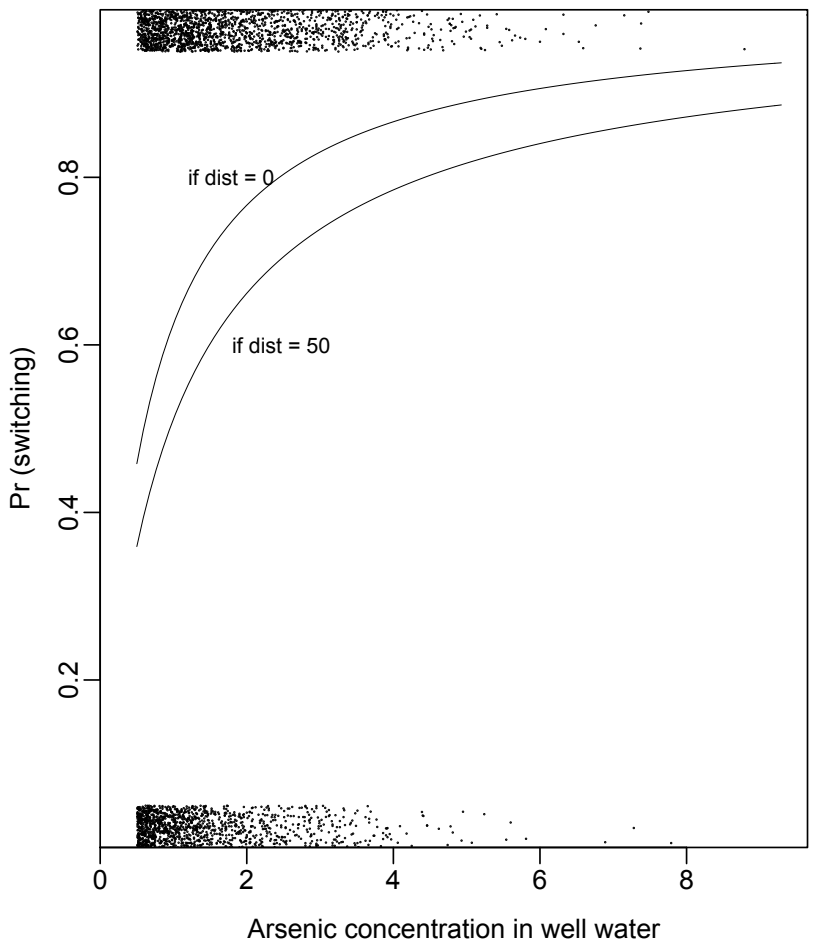


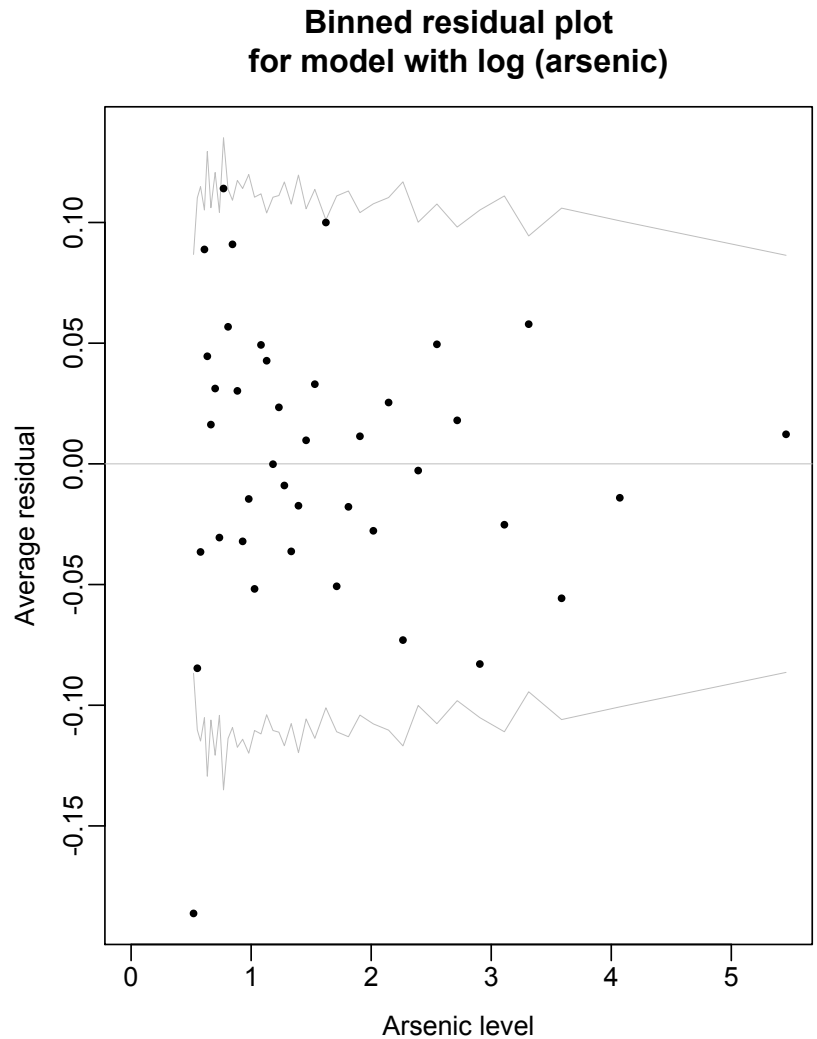


Logarithm Scale

```
> summary(fit.9a)

Call:
glm(formula = switch ~ dist100 + log.arsenic + educ4 + dist100:log.arsenic +
    dist100:educ4 + log.arsenic:educ4, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1034  -1.1623   0.7178   1.0400   1.9229

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.515987   0.103854   4.968 6.75e-07 ***
dist100            -1.338893   0.199302  -6.718 1.84e-11 ***
log.arsenic         0.906741   0.139211   6.513 7.35e-11 ***
educ4              -0.003929   0.061647  -0.064  0.94918
dist100:log.arsenic -0.156704   0.185150  -0.846  0.39735
dist100:educ4       0.338427   0.107756   3.141  0.00169 **
log.arsenic:educ4   0.060106   0.070303   0.855  0.39257
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4118.1  on 3019  degrees of freedom
Residual deviance: 3863.1  on 3013  degrees of freedom
AIC: 3877.1

Number of Fisher Scoring iterations: 4
```

Logarithm Plot





The error rate

In general

```
error.rate <- mean((predicted>0.5 & y==0) |
                   (predicted<0.5 & y==1))
```

In our model

```
> error.rate <- mean((pred.9>0.5 & switch==0) | (pred.9<0.5 & switch==1))
> error.rate
[1] 0.3649007
```

The Deviance

- Deviance is a measure of error; lower deviance means better fit to data
- If a predictor that is simply random noise is added to a model, the deviance decreases by 1, on average.
- When an informative predictor is added to a model, the deviance decreases by more than 1

The Deviance in our data set

Model	Null Deviance	Residual Deviance
dist	4118.1	4076.2
dist100	4118.1	4076.2
arsenic	4118.1	3930.7
interaction	4118.1	3927.6
center	4118.1	3927.6
social	4118.1	3905.4
educ	4118.1	3907.9
log	4118.1	3863.1

Comparing Nested Models

In linear regression: F-test for models comparisons.
In logistic regression: the deviance

$$D = -2\log(L(\hat{\beta})) = -2\log(\text{Maximum likelihood of the model})$$

Suppose model M_0 is nested within model M_1

$$D_0 - D_1 = 2\log\left(\frac{\text{Maximum likelihood of } M_0}{\text{Maximum likelihood of } M_1}\right)$$

Under null hypothesis: $D_0 - D_1 \sim \chi^2_{p-k}$

P-value ≤ 0.05: M_1 fits better than M_0

Likelihood ratio test

```
> lrtest(fit.1,fit.3)
Likelihood ratio test

Model 1: switch ~ dist
Model 2: switch ~ dist100 + arsenic
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   2 -2038.1
2   3 -1965.3  1 145.57  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Likelihood ratio test

Would you prefer the model with the log or without the log?

a) with the log

b) without the log

```
> lrtest(fit.8,fit.9)
Likelihood ratio test

Model 1: switch ~ c.dist100 + c.arsenic + c.educ4 + c.dist100:c.arsenic +
    c.dist100:c.educ4 + c.arsenic:c.educ4
Model 2: switch ~ c.dist100 + c.log.arsenic + c.educ4 + c.dist100:c.log.arsenic +
    c.dist100:c.educ4 + c.log.arsenic:c.educ4
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   7 -1945.9
2   7 -1931.5  0 28.635  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```