DSC 152:
Applied Statistical Data Analysis and Inference

Lecture #16
Introduction to Logistic Regression

Thursday, May 21
Spring Quarter 2026
Peter Chi

Binary Outcome with Quantitative Covariate(s)

Example: Diabetes and BMI

In this dataset, we have:

- An outcome variable of whether the individual gets diabetes
- A primary covariate of BMI

Our question of interest is to investigate the relationship between BMI and getting diabetes.

If you want to load the data into R yourself now, you can (but do not have to yet; we'll use it for two Your Turns later):

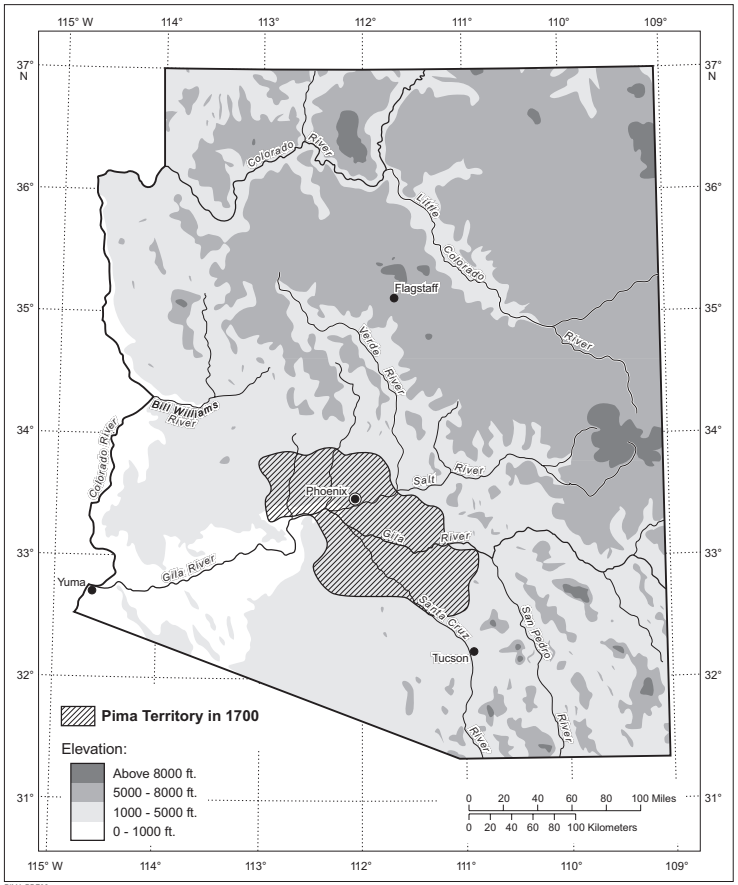
```
library(readr)
diabetes <- read_csv(
  "https://raw.githubusercontent.com/plotly/datasets/master/diabetes.csv")
```

(that can obv all go in one line; I was just trying not to make it too tiny on the slide)

Example: Diabetes and BMI

Smith et al., 1988 (paper linked on course website)

These data originally come from a study on Native Americans once known as the Pima Indian Tribe, but are now more commonly referred to as The Akimel O'odham, who are at high risk of diabetes.



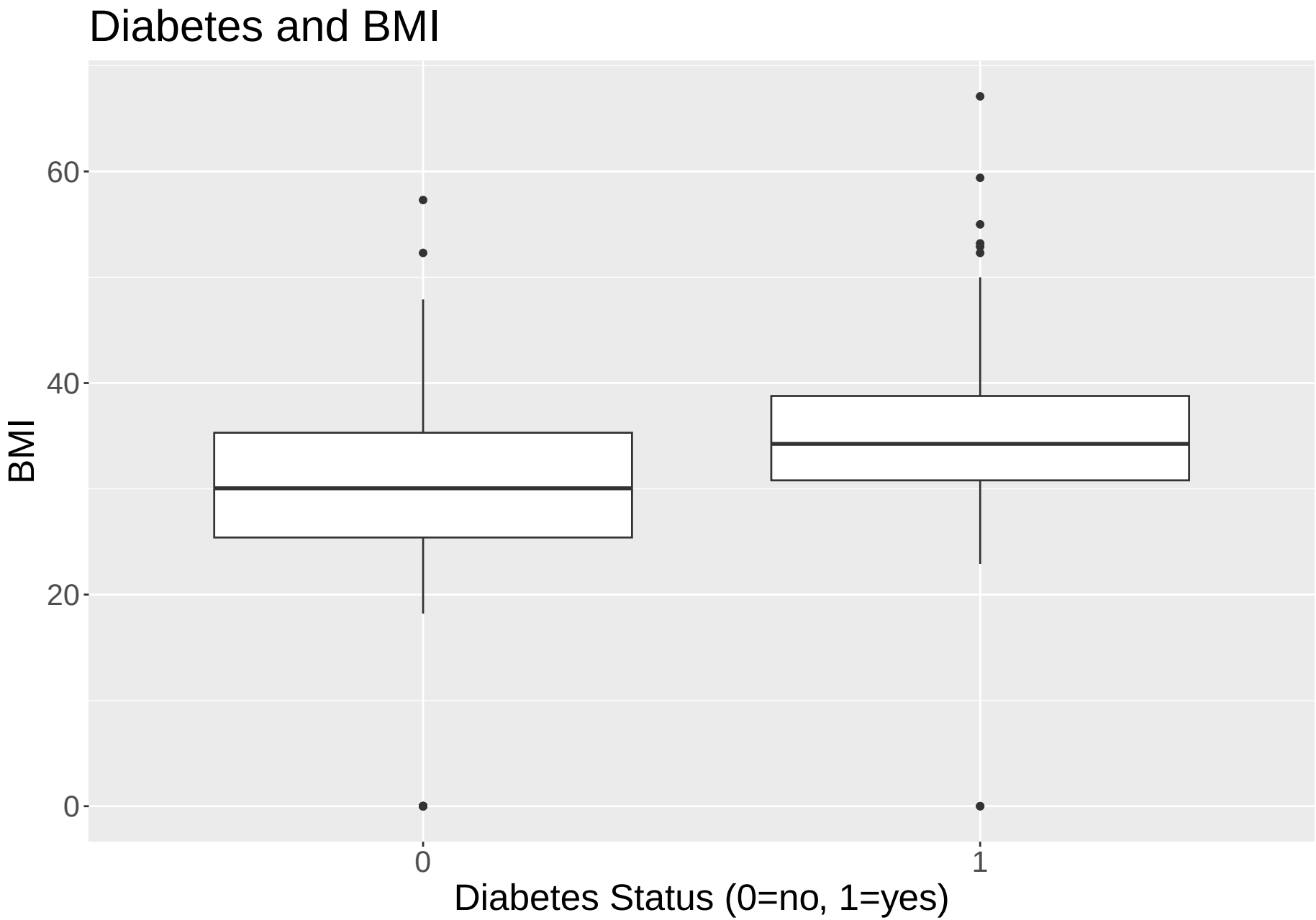
Pima Territory, 1700

ARIZONA GEOGRAPHIC ALLIANCE

Source for area of Pima Territory:
Ortiz, *Handbook*: 50

Courtesy: Arizona Geographic Alliance
Department of Geography, Arizona State University
Douglas Minnis/Barbara Trapido-Lurie

Example: Diabetes and BMI



Example: Diabetes and BMI

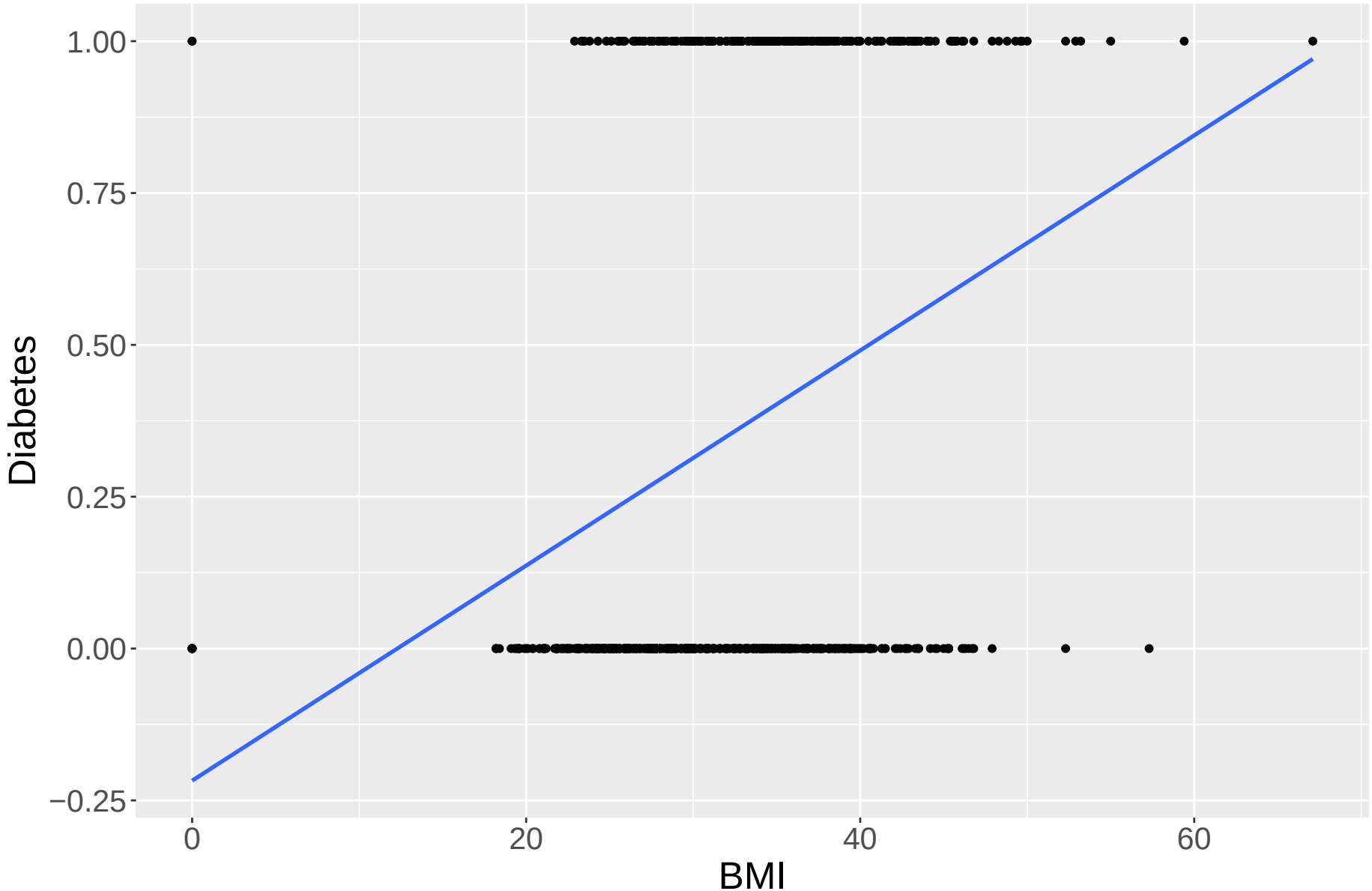
Now what?

The plot on the previous slide suggests that we have two groups of a quantitative variable, so perhaps we should do a two-sample t-Test. But this would be incorrect!!

Why?

Example: Diabetes and BMI

Ok if Diabetes is the outcome variable, then...



Example: Diabetes and BMI

So that graph and regression line is obviously stupid. On the other hand, it DOES seem to indicate that there might be a relationship...

```
##
## Call:
## lm(formula = Outcome ~ BMI, data = diabetes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7971 -0.3579 -0.2278  0.5451  1.2175
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.21752    0.06886  -3.159  0.00165 **
## BMI          0.01771    0.00209   8.472  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4564 on 766 degrees of freedom
## Multiple R-squared:  0.08567,    Adjusted R-squared:  0.08448
## F-statistic: 71.77 on 1 and 766 DF,  p-value: < 2.2e-16
```

Example: Diabetes and BMI

But we can (and should) do better than this.

What exactly should we do instead?

Last time, we learned about transformations. Can we do a transformation that would help us here?

First note: from the simple linear regression output on the previous slide, we have:

$$\hat{y} = -0.2175 + 0.0177 \cdot BMI$$

What even does $\hat{y}$ mean in this context?

Example: Diabetes and BMI

So whatever transformation we are going to do, we need it to be such that $\hat{y}$ will always be between 0 and 1, no matter what the values of the covariates are.

logit transformation

It turns out that this works quite well:

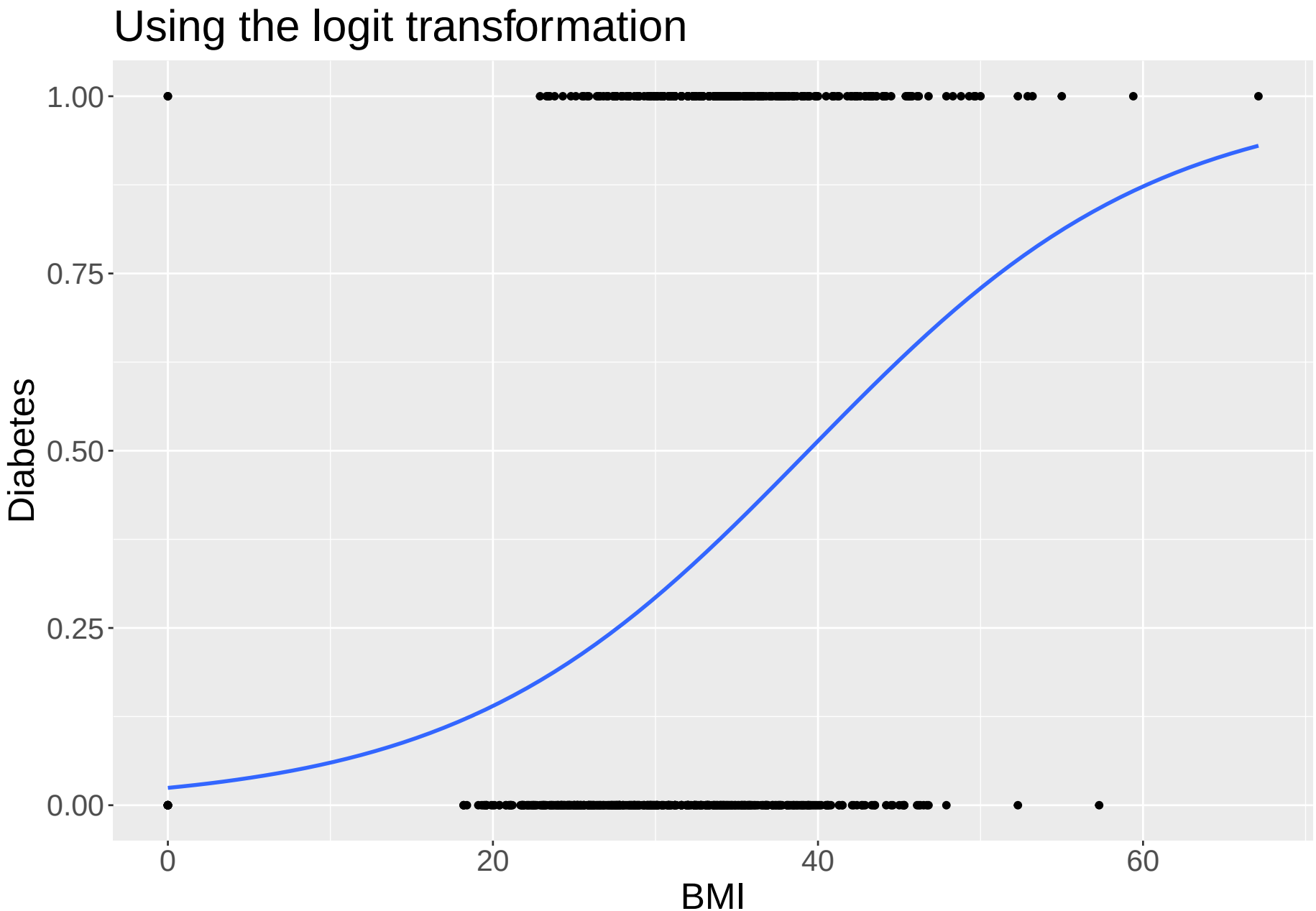
$$log\left(\frac{\hat{y}}{1-\hat{y}}\right) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$$

Wait why?

logit transformation

More space for notes if needed

Example: Diabetes and BMI



Example: Diabetes and BMI

Now, how do we fit the model? In other words, how did R get the curve on the previous slide?

Slight technical detail

Unlike in Lecture #14, we are not transforming the actual data values. *What would happen if we applied the logit transformation on the actual values of y here?*

Instead what we are doing is transforming the *predicted* value of y (which is why it is $\hat{y}$ instead of y that is shown on Slide 9).

What this means for us is that instead of using the `lm` function on the transformed data values (as we did in the last lecture), we will use the `glm` function on the original data values...

Example: Diabetes and BMI

```r
model_logit <- glm(Outcome ~ BMI, data=diabetes, family="binomial")
summary(model_logit)
```

```
##
## Call:
## glm(formula = Outcome ~ BMI, family = "binomial", data = diabetes)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.68641    0.40896  -9.014  < 2e-16 ***
## BMI          0.09353    0.01205   7.761 8.45e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 920.71  on 766  degrees of freedom
## AIC: 924.71
##
## Number of Fisher Scoring iterations: 4
```

Example: Diabetes and BMI

So this gives us the following linear model:

$$log\left(\frac{\hat{y}}{1-\hat{y}}\right) = -3.6864 + 0.0935 \cdot BMI$$

where $\hat{y}$ is the estimated probability of getting diabetes.

What are the interpretations of these coefficients??

- It is technically correct to say, e.g., that
  - 0.0935 is the estimated difference in $log\left(\frac{\hat{y}}{1-\hat{y}}\right)$ corresponding to a 1-unit increase in BMI.
- But can we get a more meaningful interpretation than that?
- Note the familiarity of the quantity $\frac{\hat{y}}{1-\hat{y}}$ as we just learned about odds ratios...

Example: Diabetes and BMI

Logistic Regression and Odds Ratios

Here's the model again:

$$log\left(\frac{\hat{y}}{1-\hat{y}}\right) = -3.6864 + 0.0935 \cdot BMI$$

which is also equivalent to:

$$\begin{aligned}\frac{\hat{y}}{1-\hat{y}} &= e^{-3.6864+0.0935 \cdot BMI} \\ &= \left(e^{-3.6864}\right)\left(e^{0.0935}\right)^{BMI}\end{aligned}$$

What does a 1-unit increase in BMI do to $\hat{y}$ now?

Example: Diabetes and BMI

More space for notes if needed

Conclusions

What p-value(s) from the output do we care about?

Recall that our question of interest is in whether BMI is associated with getting diabetes, and our logistic regression model is:

$$log\left(\frac{\widehat{diabetes}}{1-\widehat{diabetes}}\right) = \hat{\beta}_0 + \hat{\beta}_1 \cdot BMI$$

which gives us the following hypotheses:

$$H_0 \colon \beta_1 = 0$$
$$H_A \colon \beta_1 \neq 0$$

So the p-value for this is... pollev.com

Conclusions

How about a confidence interval?

```
confint(model_logit)
```

```
## Waiting for profiling to be done...
```

```
##                   2.5 %     97.5 %
## (Intercept) -4.50912343 -2.9051697
## BMI          0.07045954  0.1177311
```

But now on the odds ratio scale...

```
exp(confint(model_logit)[2,])
```

```
## Waiting for profiling to be done...
```

```
##    2.5 %   97.5 %
## 1.073001 1.124942
```

Conclusions

A full written conclusion

We find that there is statistically significant evidence at the $\alpha = 0.05$ level that BMI is associated with getting diabetes (p=8.45×10^{-15}). Specifically, we estimate that a 1-unit increase in BMI corresponds to an odds ratio of 1.098 for diabetes, with 95% CI: (1.073, 1.1249).

Conclusions

A full written conclusion

We find that there is statistically significant evidence at the $\alpha = 0.05$ level that BMI is associated with getting diabetes (p=8.45×10^{-15}). Specifically, we estimate that a 1-unit increase in BMI corresponds to an odds ratio of 1.098 for diabetes, with 95% CI: (1.073, 1.1249).

Note: the text above was written with inline R code

```
We find that there is statistically significant evidence at the
$\alpha=0.05$ level that BMI is associated with getting diabetes (p=`r
signif(summary(model_logit)$coefficients[2,4], 4)`). Specifically, we
estimate that a 1-unit increase in BMI corresponds to an odds ratio of `r
round(exp(summary(model_logit)$coefficients[2,1]), 4)` for diabetes, with
95\% CI: (`r round(exp(confint(model_logit, 'BMI')[1]), 4)`, `r
round(exp(confint(model_logit, 'BMI')[2]), 4)`).
```

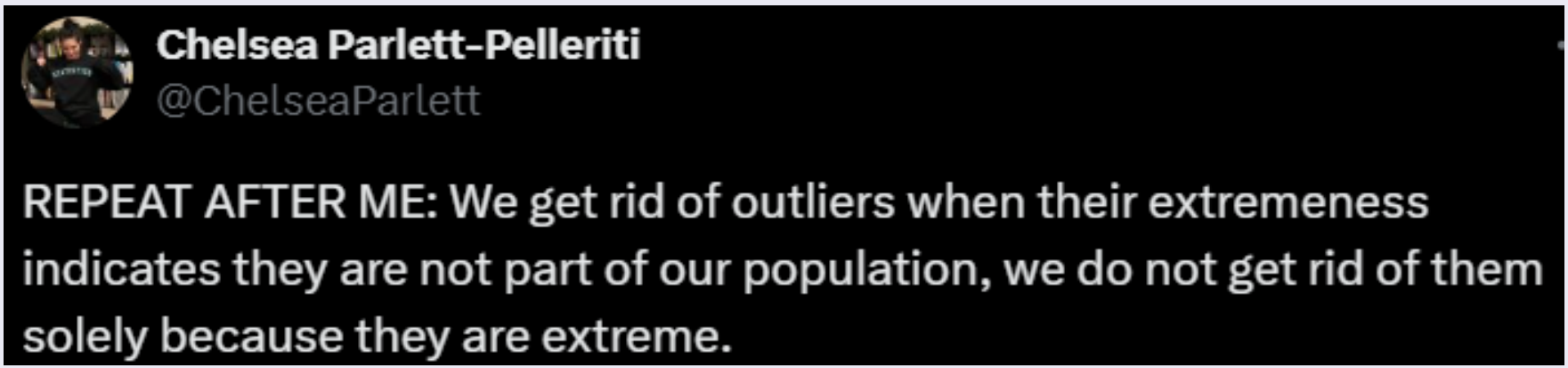
Your Turn #1

Quote of the Day #1

"The data are what they are."

Dr. Rich Hume
Department of Biology
University of Michigan

Quote of the Day #2



Your Turn #1

There are some values of BMI in the dataset that are impossible. First, load the dataset in now if you haven't already. Then,

- Remove the rows with impossible values of BMI, and then re-run the primary analysis.

- Report your conclusions at $\alpha = 0.05$ for the primary question of interest, along with an estimated odds ratio and 95% confidence interval.

Conditions for Logistic Regression

- The outcome variable is binary
- The observations are independent of each other
- The relationship between your covariates and the log-odds of the outcome variable is linear

Conditions for Logistic Regression

- The outcome variable is binary
  - This one is obvious
- The observations are independent of each other
  - We could check residuals against time or something like that, but also can just make arguments based on anything we know about how the data were collected
- The relationship between your covariates and the log-odds of the outcome variable is linear
  - We can check this one (somewhat) similar to how we checked for linearity in linear regression...

Conditions for Logistic Regression

Linearity Condition

Can we just check residuals vs. fitted values?

```
diabetes$fitted <- model_logit$fitted.values
diabetes$residuals <- residuals(model_logit, type="deviance")
ggplot(data=diabetes, aes(x=fitted, y=residuals)) +
  geom_point()
```

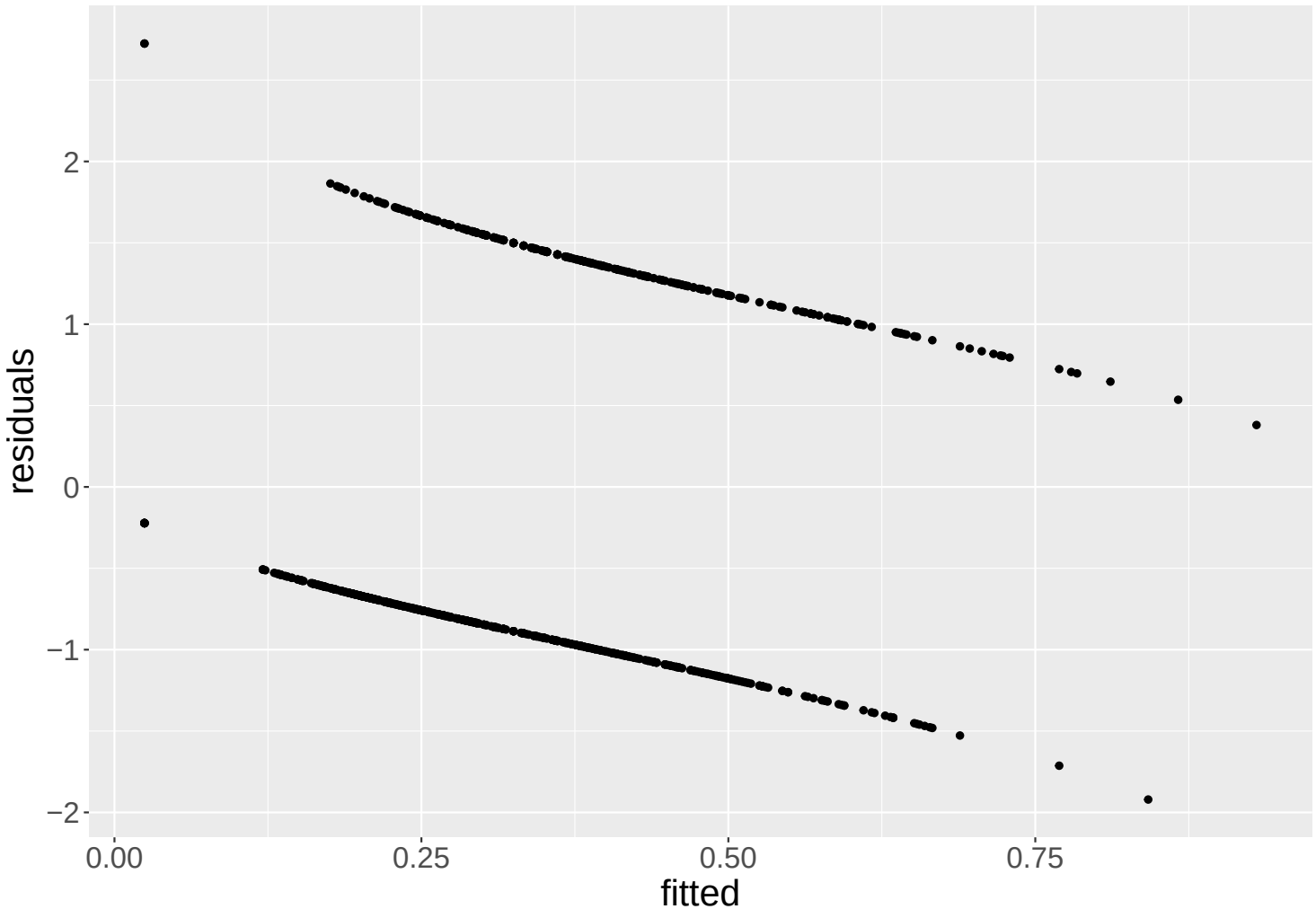
What's up with the residuals(model_logit, type="deviance")?

When using `glm` for logistic regression, it is customary to calculate "deviance residuals" instead of grabbing `residuals` from the model output. For our purposes, all you need to know is that they have better properties in a glm, and that you can get them with the above code.

Conditions for Logistic Regression

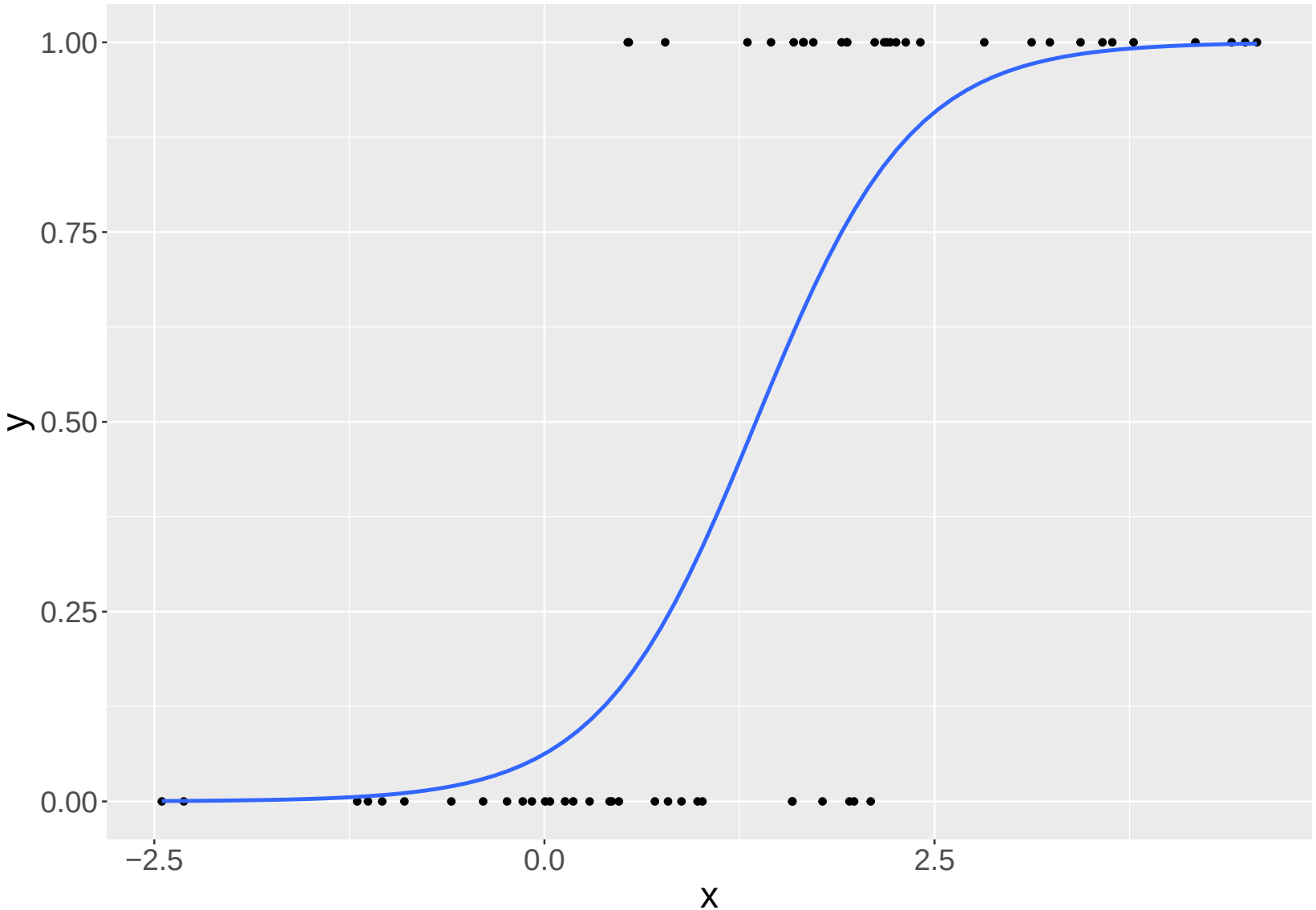
Linearity Condition

What exactly is happening here?



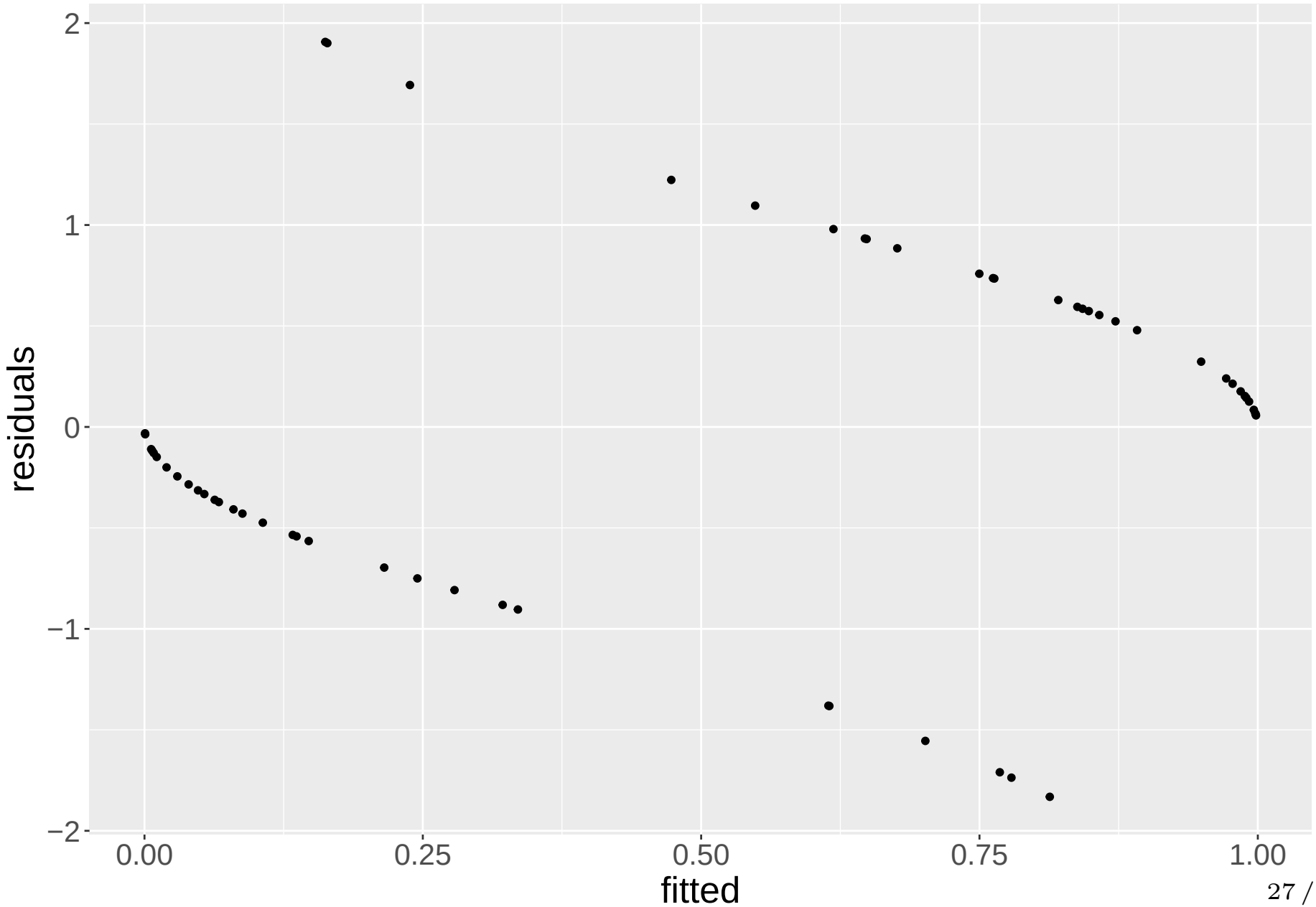
Conditions for Logistic Regression

Let's illuminate this with simpler examples



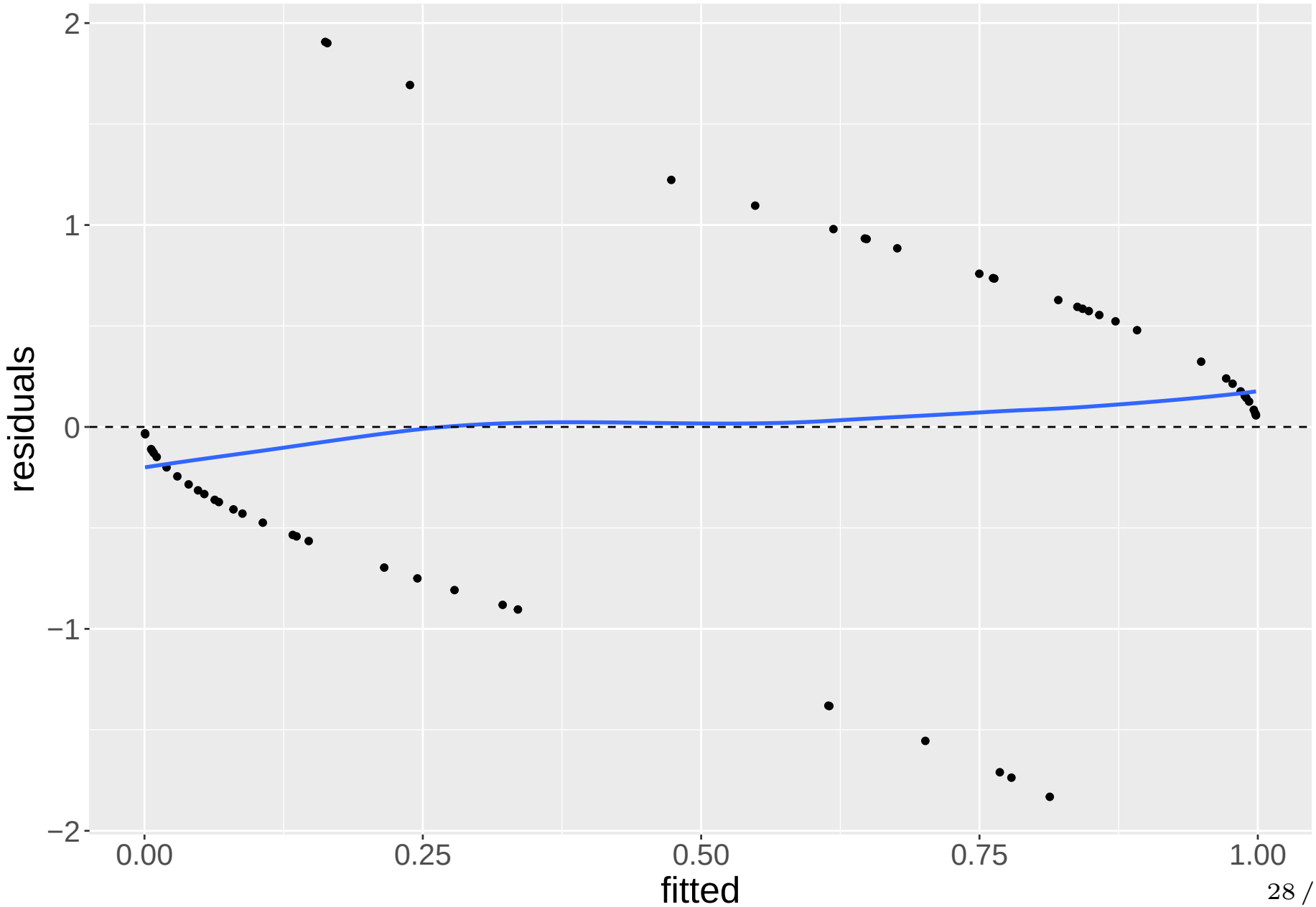
Conditions for Logistic Regression

Let's illuminate this with simpler examples



Conditions for Logistic Regression

Let's illuminate this with simpler examples



Conditions for Logistic Regression

Let's illuminate this with simpler examples

So this is good, because the we are looking for the loess smoother to roughly follow the flat horizontal line at 0.

Here is the code that made the plot on the previous slide:

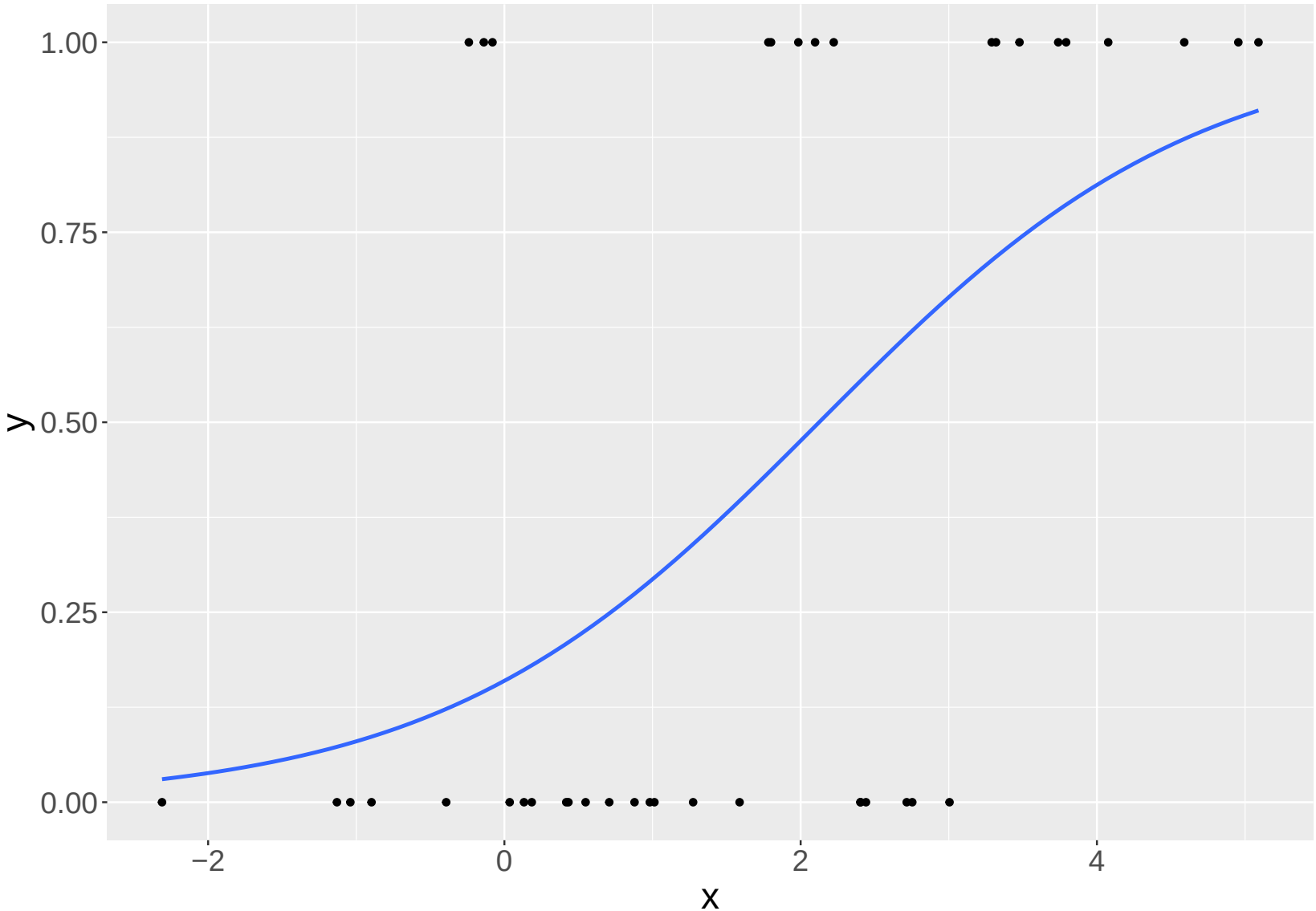
```
model_simp <- glm(y ~ x, family="binomial")

df$fitted <- model_simp$fitted.values
df$residuals <- residuals(model_simp, type="deviance")
ggplot(data=df, aes(x=fitted, y=residuals)) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE) +
  geom_hline(yintercept = 0, linetype = "dashed")
```

Conditions for Logistic Regression

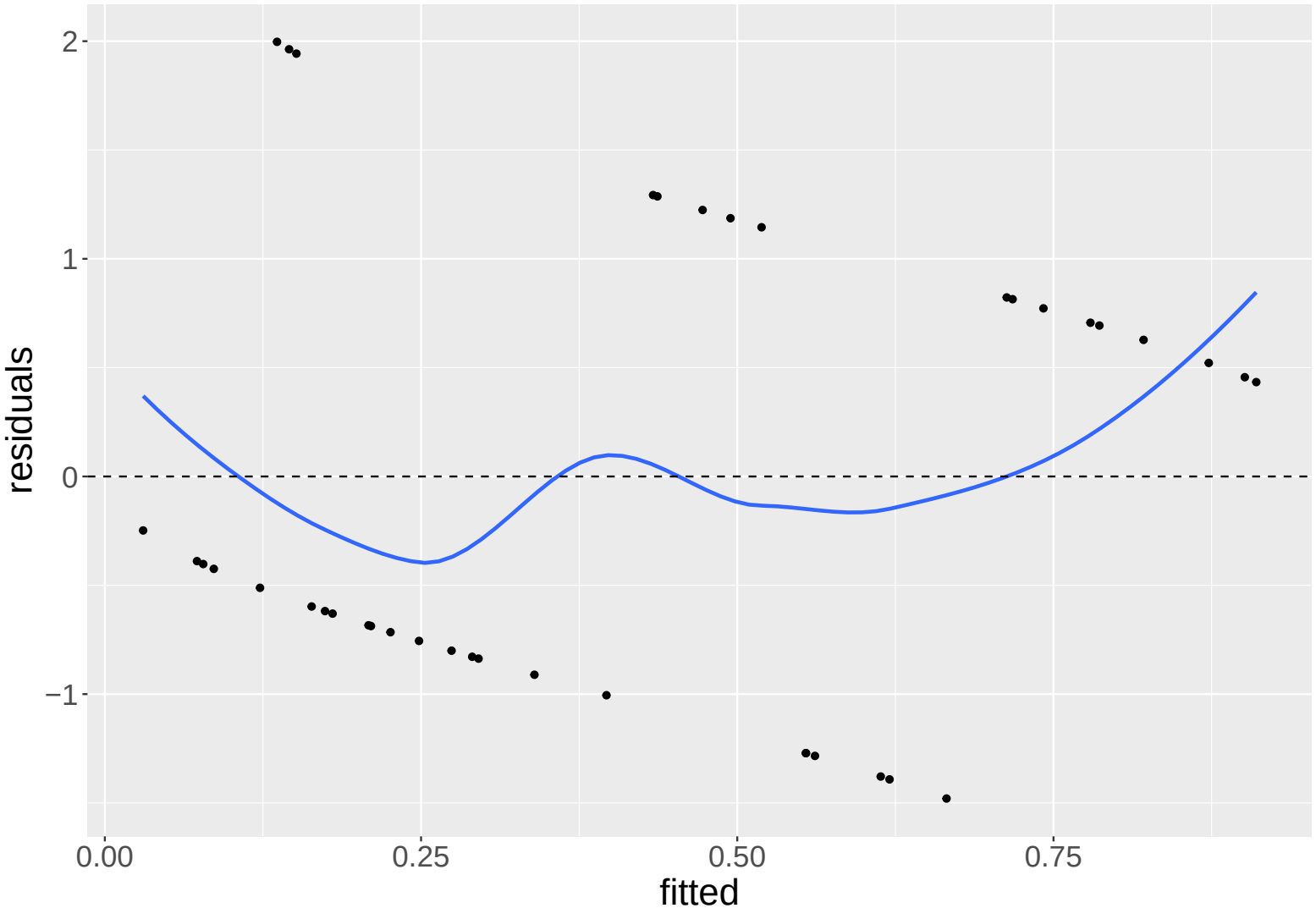
Let's illuminate this with simpler examples

Now here's an example in which we have a questionable fit:



Conditions for Logistic Regression

Let's illuminate this with simpler examples

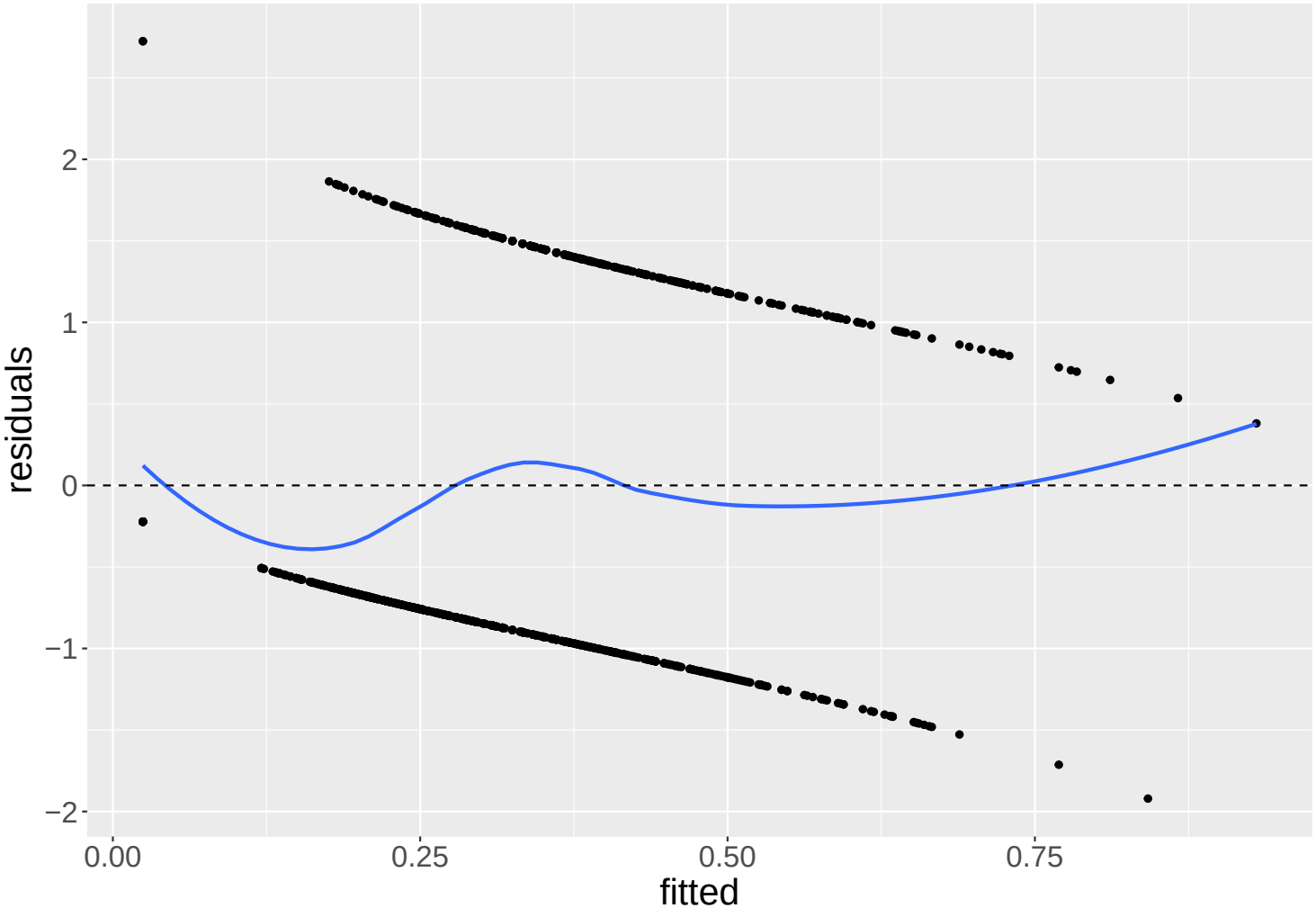


This loess curve wobbles a lot, so this one is not so good.

Conditions for Logistic Regression

Linearity Condition

Back to the BMI data:



Your Turn

Now we will consider the full `diabetes` dataset:

columns	description
Pregnancies	# of times pregnant
Glucose	Plasma glucose level
BloodPressure	Diastolic bp
SkinThickness	Triceps skin fold thickness
Insulin	2-hour serum insulin
BMI	body mass index
DiabetesPedigreeFunction	genetic disposition to diabetes
Age	age in years
Outcome	Has diabetes or not

Your Turn

Suppose that the primary analysis was the simple logistic regression model with just BMI as a covariate, but now we would like to perform a sensitivity analysis:

- Perform backward selection via p-values, starting from the model with all 8 covariates in it
  - Display the model summary at each step
- For the final model, write a brief description of the following:
  - The p-value for the primary question and what the conclusion would have been if this model had actually been the one used for inference
  - The interpretation, on the odds ratio scale, of the estimate of the primary coefficient of interest, along with a 95% confidence interval
- Check the linearity condition via a residual plot with the final model, and briefly comment

Recap and Looking Ahead

Recap

- Logistic regression is used when we have a binary outcome variable, and any number of quantitative/categorical/binary covariates
- $e^{\hat{\beta}_i}$ can be interpreted as an odds ratio

Looking Ahead

Time Series! Also next Tuesday's class is asynchronous.

Today's Daily Check

The two Your Turns