Chapter 15: Regression III

DSCC 462

Computational Introduction to Statistics

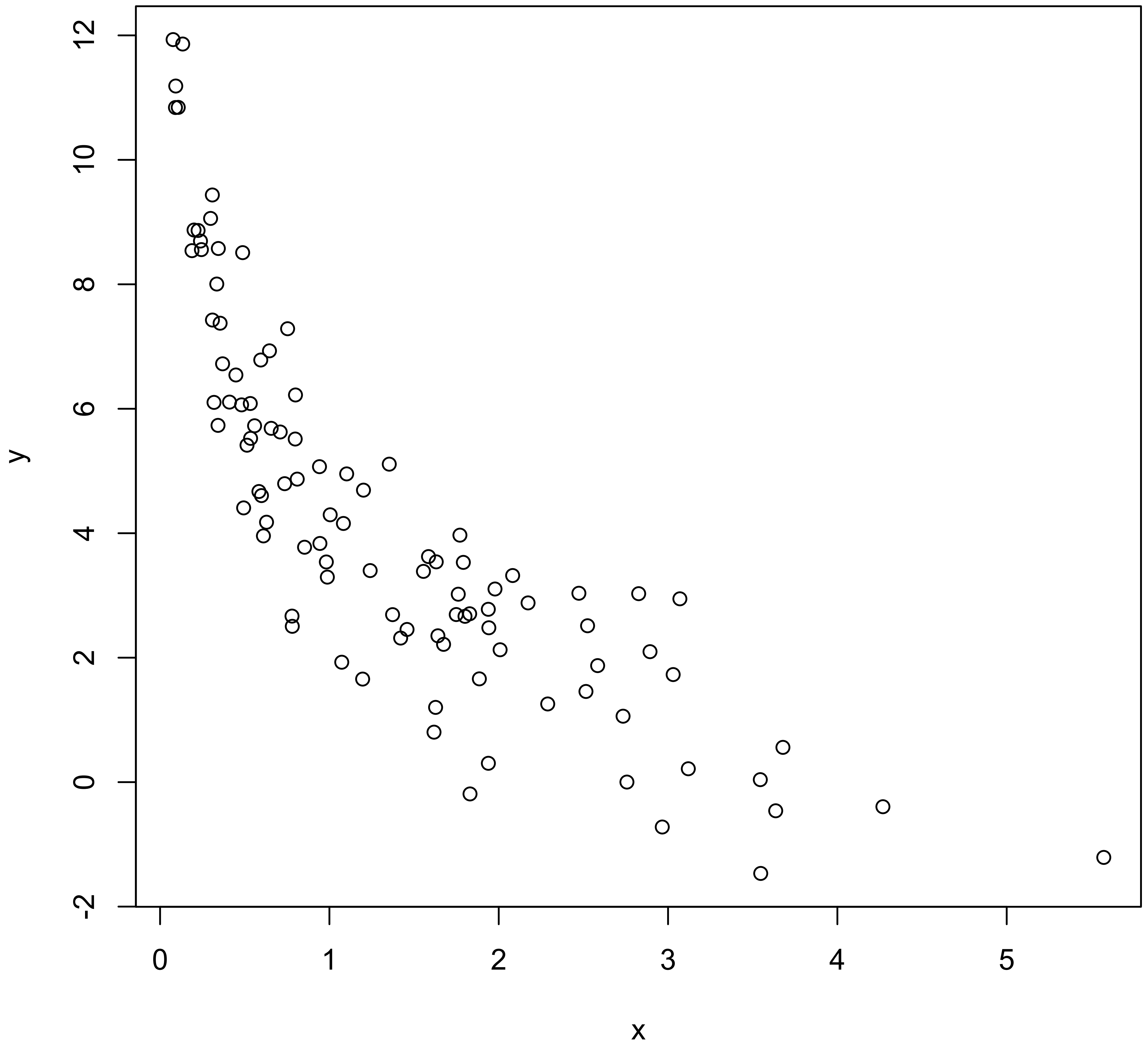
Anson Kahng

Fall 2022

Plan for Today

- Transformations of variables
- Categorical variables with multiple categories
- Predicting binary random variables (logistic regression)

Transformations



Transformations

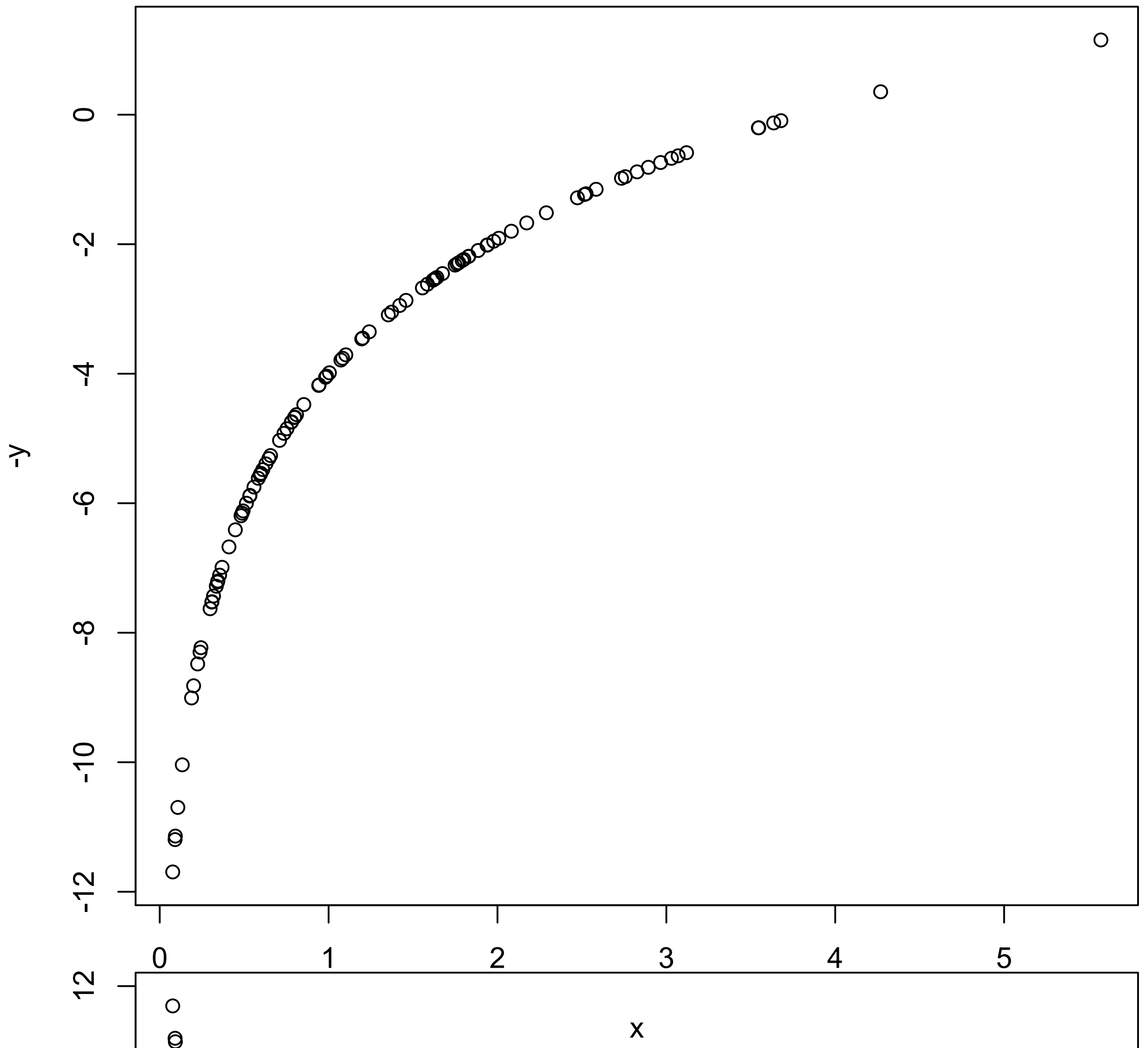
- Typically, we will apply transformations of the form x^p or y^p, for $p = \ldots, -3, -2, -1, -\frac{1}{2}, \frac{1}{2}, 1, 2, 3, \ldots$
- Or, we will use the natural log: $\ln(x)$ or $\ln(y)$ – corresponds to a choice of $p = 0$ in the above power transformations
- To determine which transformation is a good place to start, we use the (*Tukey) ladder of powers*

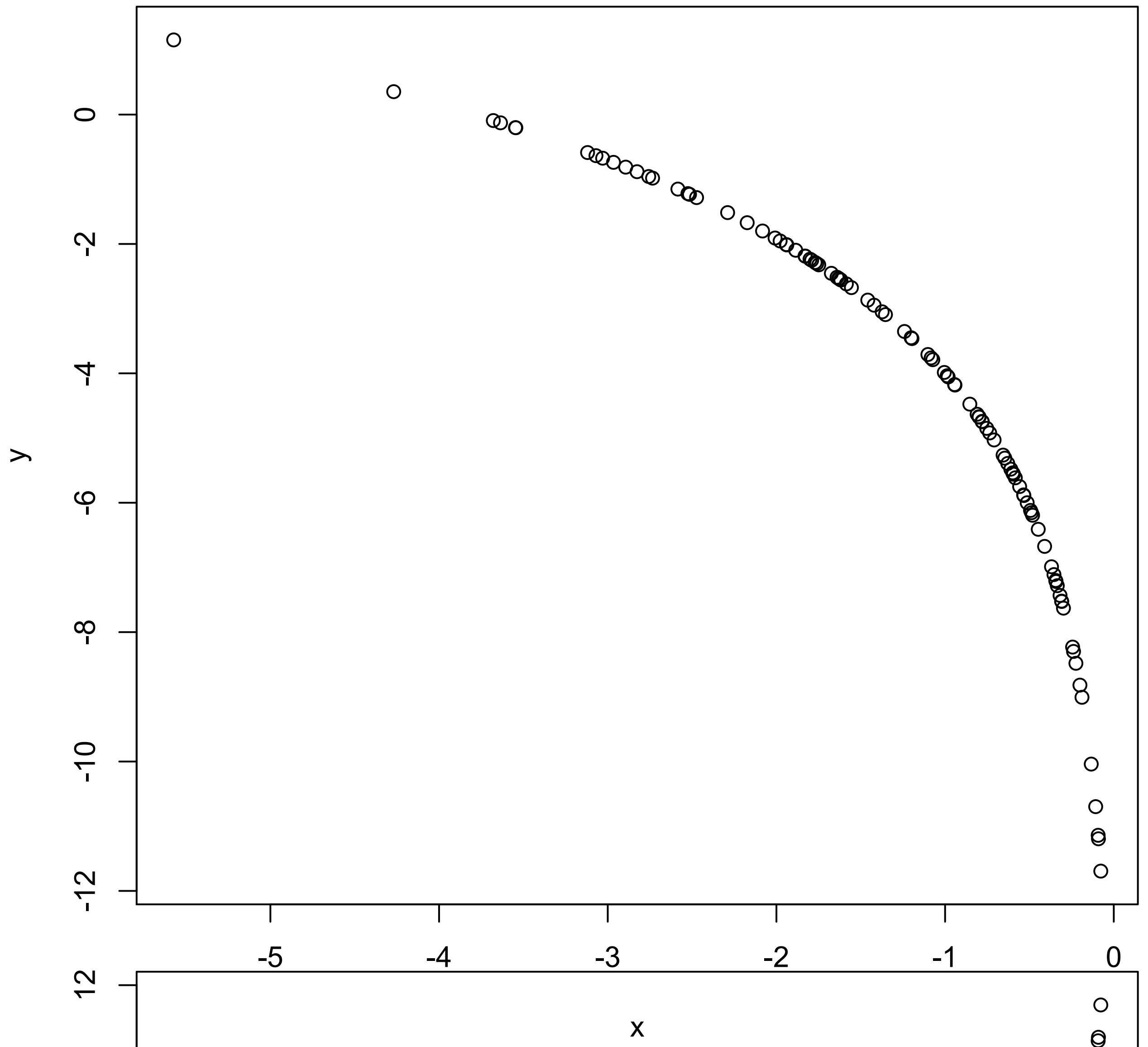
Transformations

- As originally described by Tukey:
  - Visually divide your data into even thirds, take median x and y value in each third to get three reference points
  - Draw lines connecting consecutive reference points
  - Draw an arrow toward the “elbow” of the lines, then use this direction to decide how to transform data

Transformations

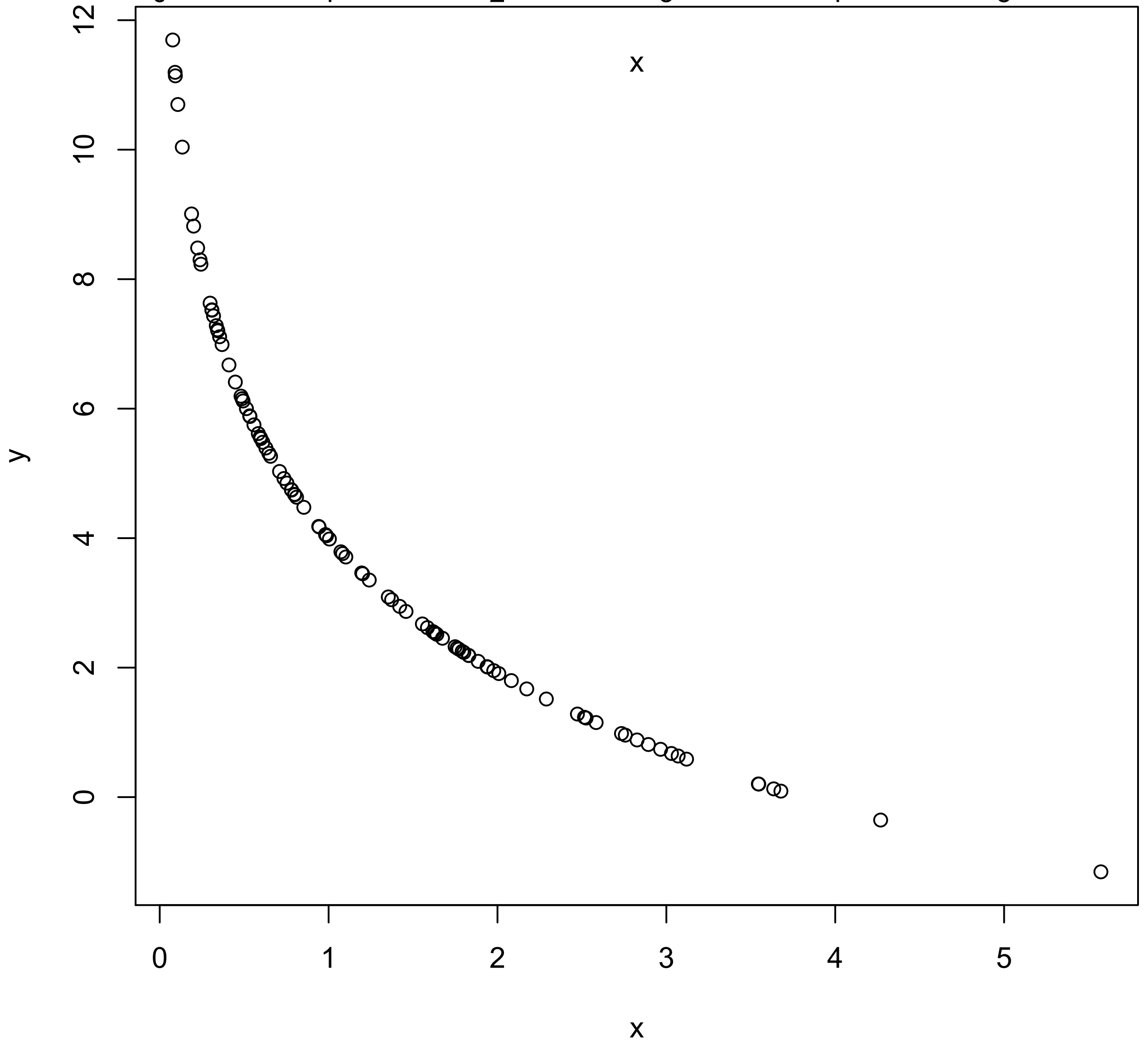
Decrease exponent of x
Increase exponent of y

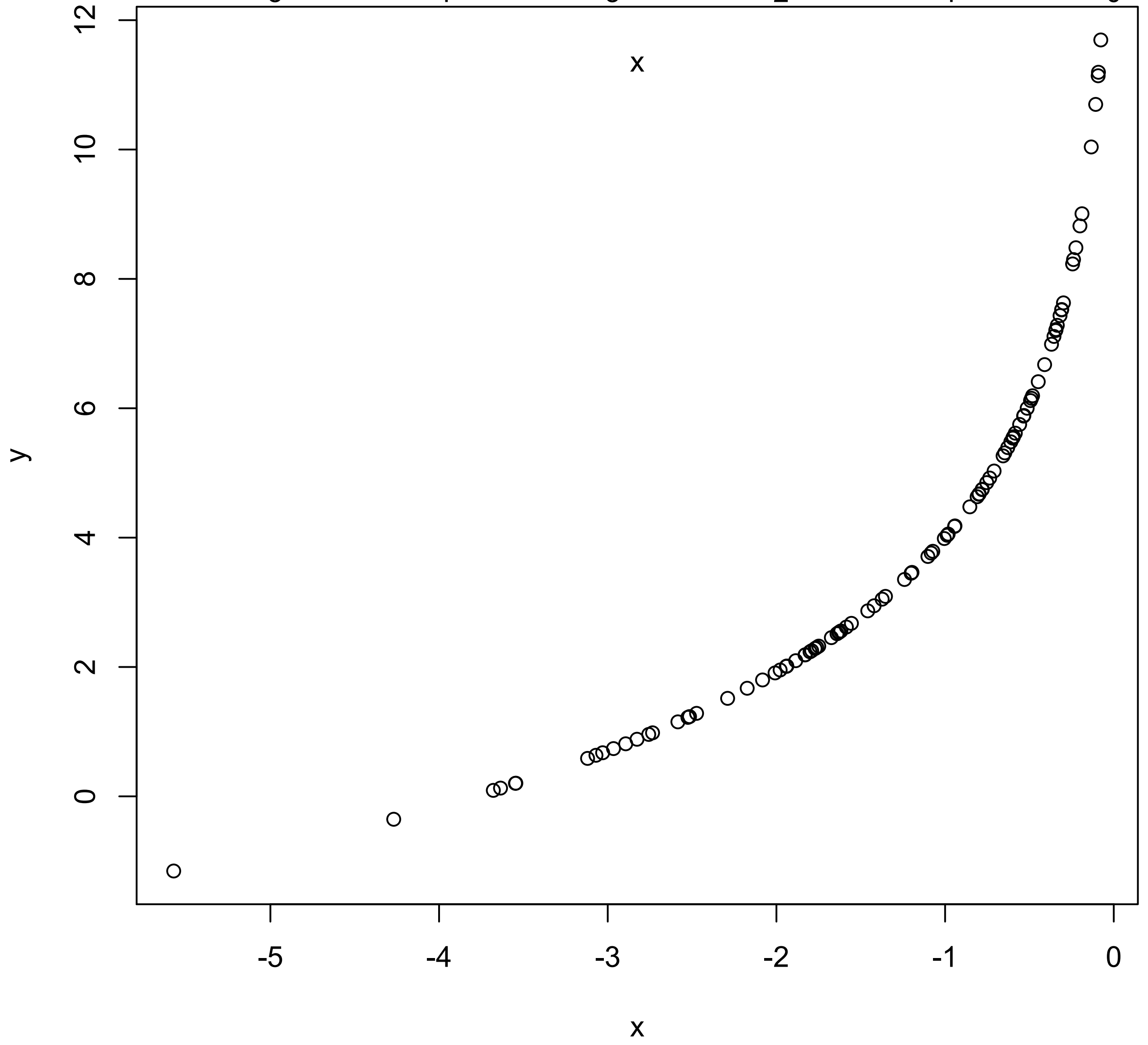




Increase exponent of x
Increase exponent of y

Decrease exponent of x
Decrease exponent of y

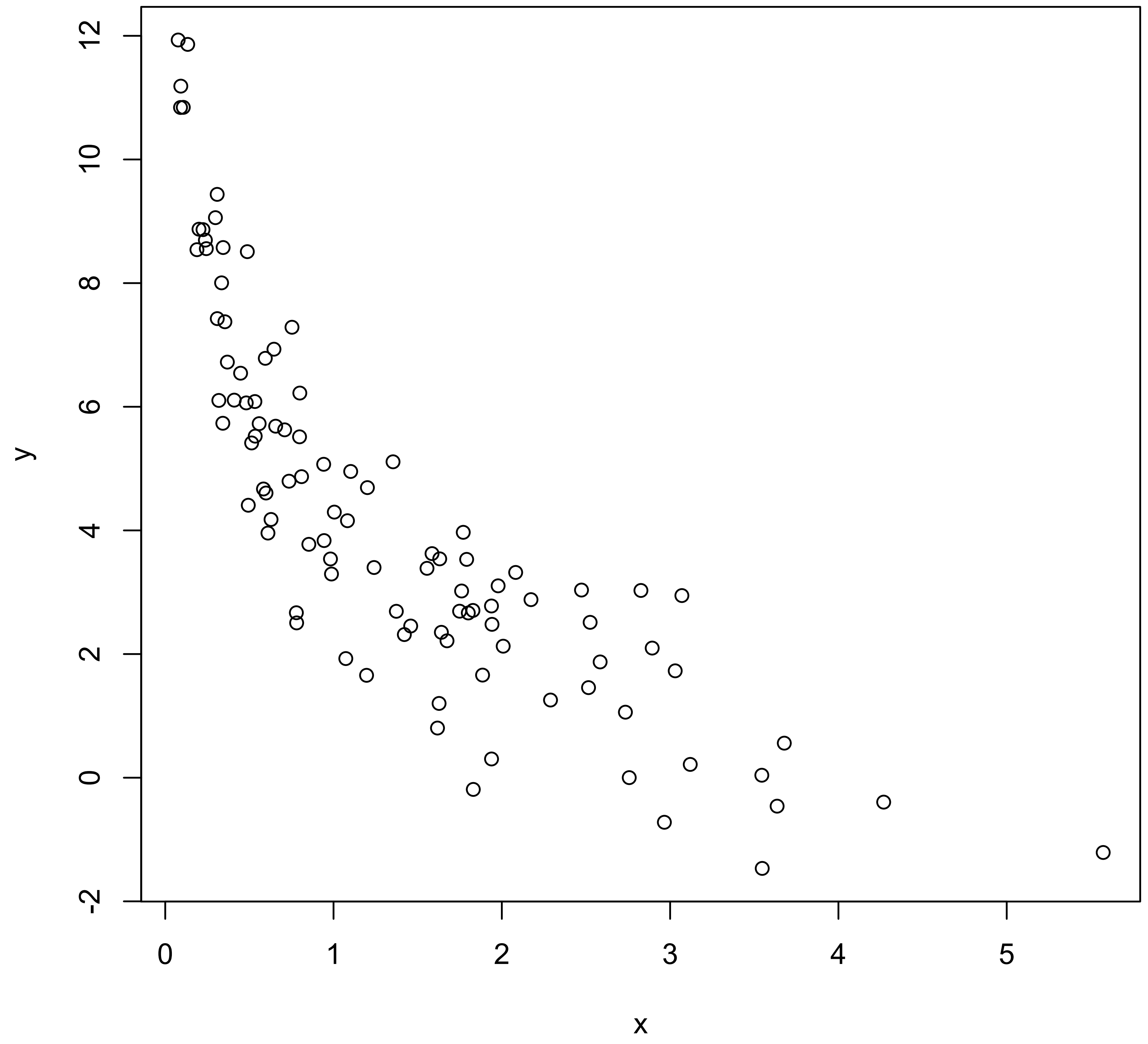




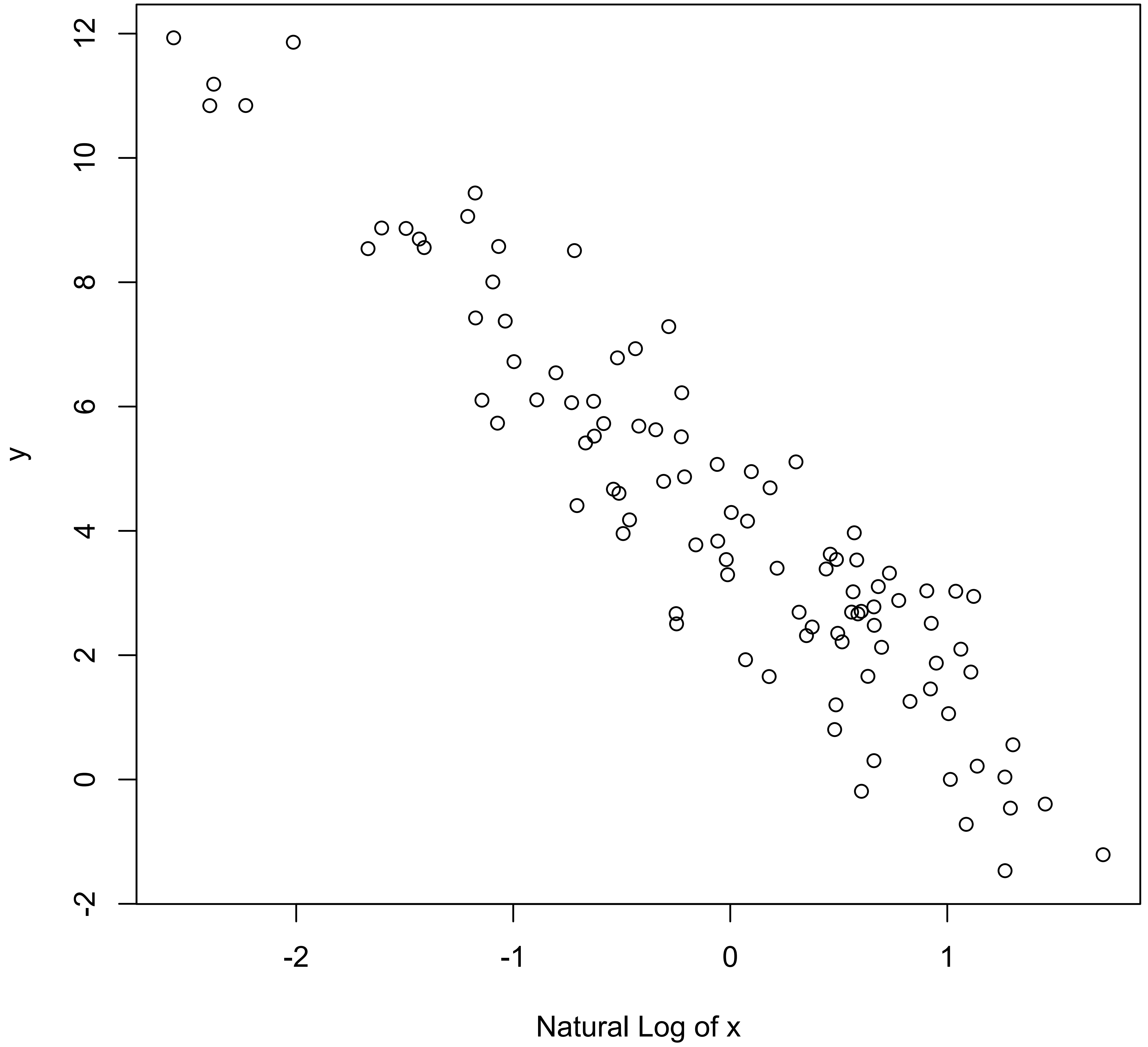
Increase exponent of x
Decrease exponent of y

Transformations

- Consider the following plot
- A natural log transformation of x may be appropriate, since we need to decrease the exponent



Transformed Data



Interpretation of Logarithmic Variables

- Consider a setting with a log-transformed x variable: $\hat{y} = \hat{\beta}_1 \ln x + \hat{\beta}_0$
- How can we interpret $\hat{\beta}_1$?
- For linear x variables, $\hat{\beta}_1$ is the increase in $\hat{y}$ associated with a one-unit increase in x
  - For logarithmic x variables, it's a bit different
- Compare $\hat{y} = \hat{\beta}_1 \ln x + \hat{\beta}_0$ with $\hat{y}* = \hat{\beta}_1(\ln x + 1) + \hat{\beta}_0$
  - Adding 1 to $\ln x$ is equivalent to multiplying x by e
  - $\hat{\beta}_1$ is the increase in $\hat{y}$ associated with multiplying x by $e \approx 2.718$

Interpretation of Logarithmic Variables

- What if we want something more easily interpretable?
- Instead of adding 1 to $\ln x$ (i.e., multiplying x by e), what if we looked at a $p \cdot 100\,\%$ increase in x (i.e., multiplying x by $(1+p)$)?
- Compare $\hat{y} = \hat{\beta}_1 \ln x + \hat{\beta}_0$ with $\hat{y}* = \hat{\beta}_1 \ln(x \cdot (1+p)) + \hat{\beta}_0$
- $\hat{y}* \approx \hat{\beta}_1 \ln x + \hat{\beta}_0 + \hat{\beta}_1 p$ for small p because $\ln(1+x) \approx x$ for small positive x
- **Interpretation**: $p \cdot 100\,\%$ increase in $x \to \hat{y}$ increases by $\hat{\beta}_1 \cdot p$
  - If x increases by 1%, $\hat{y}$ increases by $\hat{\beta}_1 / 100$

Interpretation of Logarithmic Variables

- Consider a setting with a log-transformed y variable: $\ln \hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$
- Each one-unit increase in x multiplies the value of $\hat{y}$ by $e^{\hat{\beta}_1}$
- For small values of $\hat{\beta}_1$, we have $e^{\hat{\beta}_1} \approx 1 + \hat{\beta}_1$, or an increase of $\hat{\beta}_1 \cdot 100\,\%$
- **Interpretation**: If x increases by one unit, $\hat{y}$ increases by $\hat{\beta}_1 \cdot 100\,\%$

Interpretation of Logarithmic Variables

- Consider a setting with two log-transformed variables: $\ln \hat{y} = \hat{\beta}_1 \ln x + \hat{\beta}_0$
- Multiplying x by e multiplies the value of $\hat{y}$ by $e^{\hat{\beta}_1}$
- For a $p \cdot 100\,\%$ increase in x, y changes by a factor of $e^{\ln(1+p)\cdot\hat{\beta}_1} \approx 1 + p\hat{\beta}_1$ for small p
- **Interpretation**: If x increases by $p \cdot 100\,\%$, $\hat{y}$ increases by $p \cdot \hat{\beta}_1 \cdot 100\,\%$

Multiple Linear Regression: Categorical Variables

- Recall: For binary (indicator) random variables, we assign one category a value of 1 and the other category a value of 0
- What happens if we have a categorical random variable with more than two categories?
  - E.g., eye color can take values {brown, blue, hazel, amber, other}
- How can we use these variables in linear regression?

Multiple Linear Regression: Categorical Variables

- Idea 1: Assign each category a number
  - E.g., for eye color = {brown, blue, hazel, amber, other}, let brown = 0, blue = 1, hazel = 2, amber = 3, and other = 4
- However, there is a problem with this approach…
  - This implies that there is an ordering over categories, and stipulates that changing eye color from brown to amber is three times as meaningful as changing eye color from brown to blue
  - We need another approach

Multiple Linear Regression: Categorical Variables

- Idea 2: Create a new binary variable for each possible value of eye color
  - E.g., for eye color = {brown, blue, hazel, amber, other}, create five new binary variables x_{brown}, x_{blue}, x_{hazel}, x_{amber}, and x_{other}
  - For each data point, exactly one of these variables will be 1, and the rest will be 0
- Extension of “dummy coding” from previous lecture (for a single binary variable)
  - R does this by default!

Multiple Linear Regression: Categorical Variables

- Benefits of dummy coding:
  - Relatively simple to set up
  - Easy to map features to binary 0/1s
- Drawbacks:
  - Relatively primitive (there are other more expressive ways of encoding categorical variables)
  - If the variable has many categories, have to create many dummy variables

Logistic Regression

- Up to this point, we’ve looked at continuous or categorical explanatory variable(s) and a continuous response variable
  - E.g., how do age, weight, and/or eye color predict height?
- Now, what if the variable we want to predict is binary?
  - E.g., how do age, weight, and/or eye color predict diabetes?

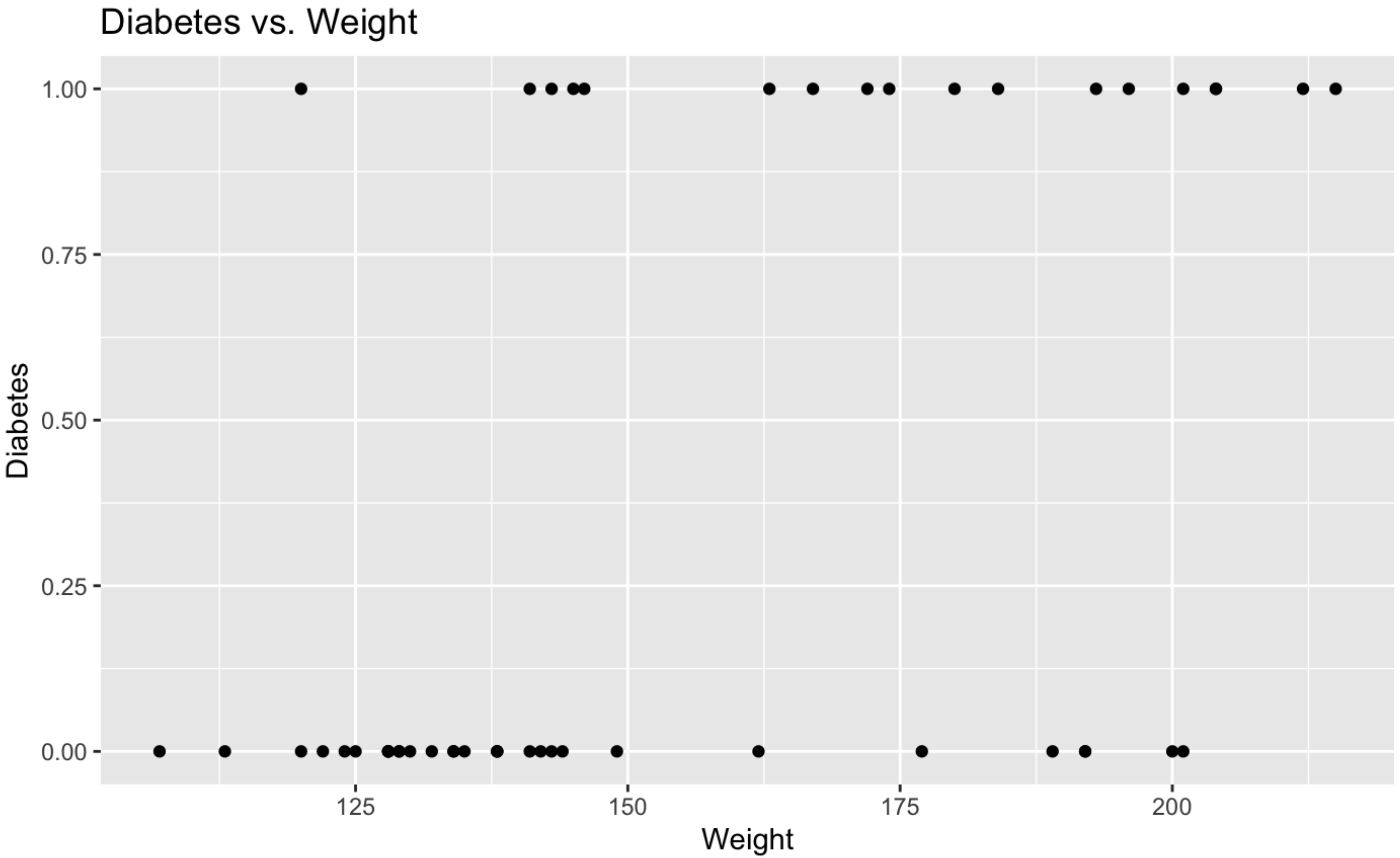
Odds

- Odds are an alternative way of expressing the probability of an event occurring
- For an event E, $\mathrm{odds}(E) = \dfrac{\Pr(E)}{\Pr(E^c)}$ (the probability of the event occurring divided by the probability of the event not occurring)
- If we are told that the odds of event E happening are x to y, then we have $\mathrm{odds}(E) = \dfrac{x}{y}$
- Extracting probabilities: $\Pr(E) = \dfrac{x}{x+y}$, and $\Pr(E^c) = \dfrac{y}{x+y}$

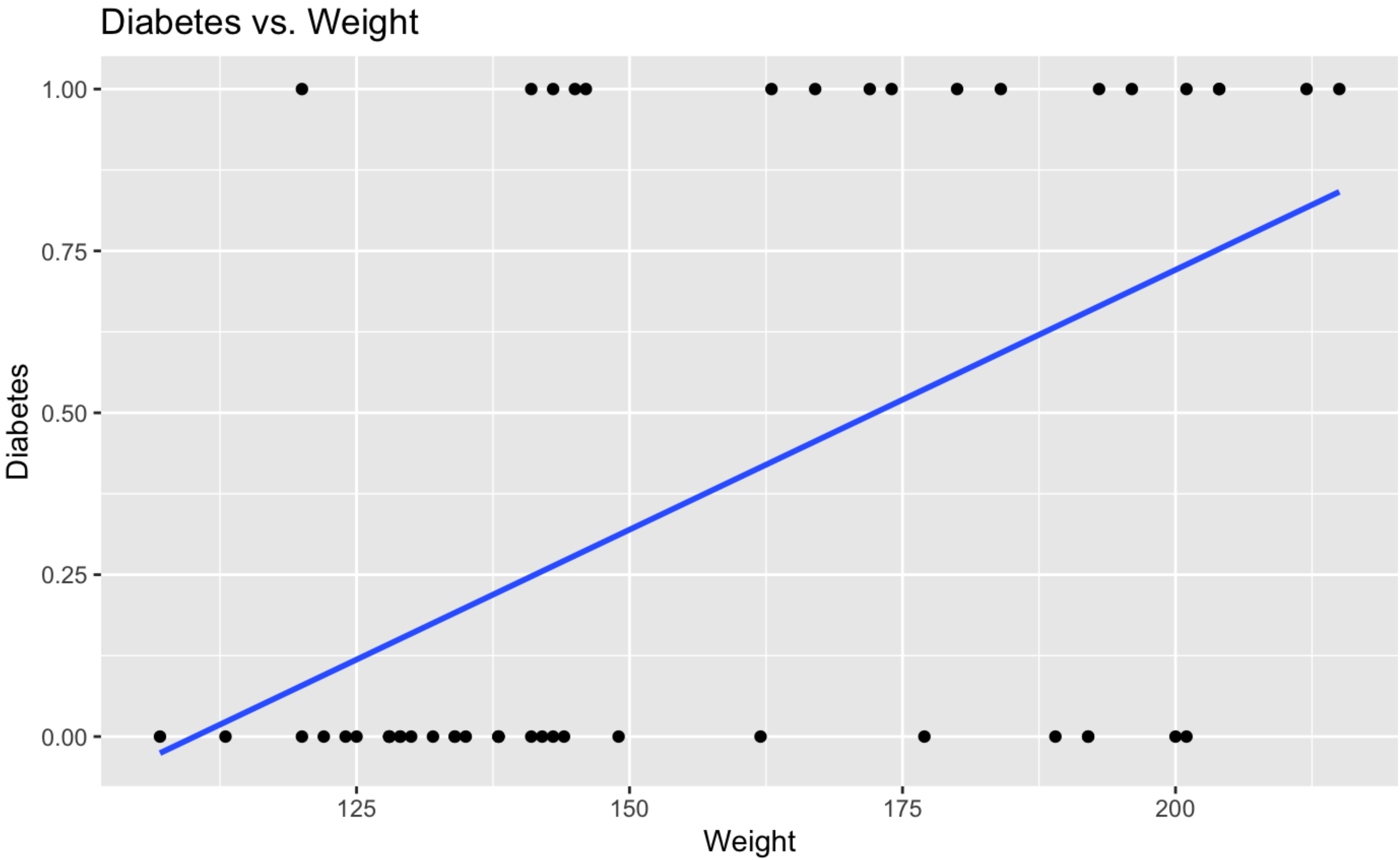
Logistic Regression: Motivating Example

- Suppose that we would like to predict whether or not someone has diabetes based on various measurements about them
- Response variable: diabetes (binary)
- Explanatory variables: weight, height, age, sex
  - For now, just consider weight for simple logistic regression

Logistic Regression: Motivating Example

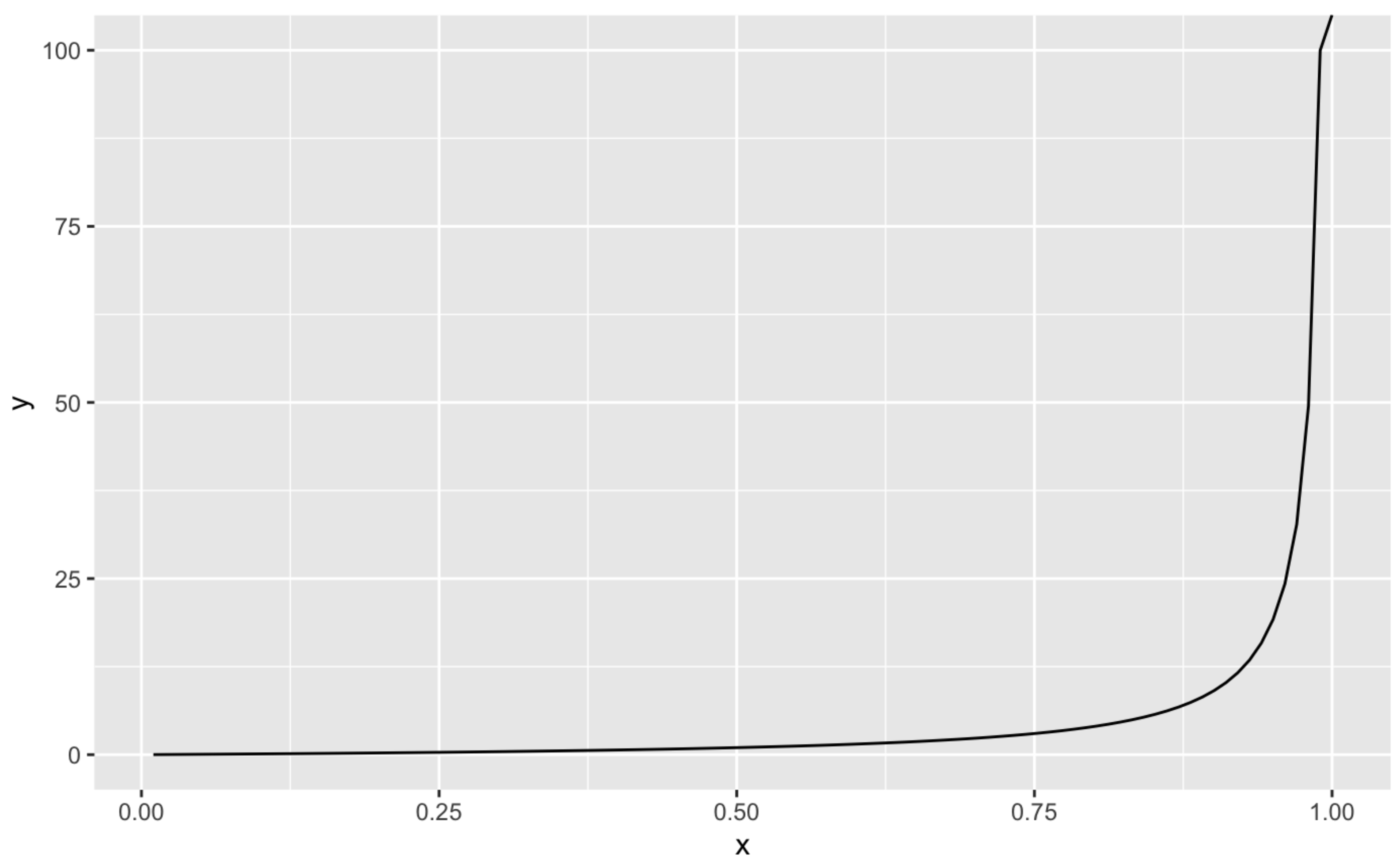


Logistic Regression: Motivating Example



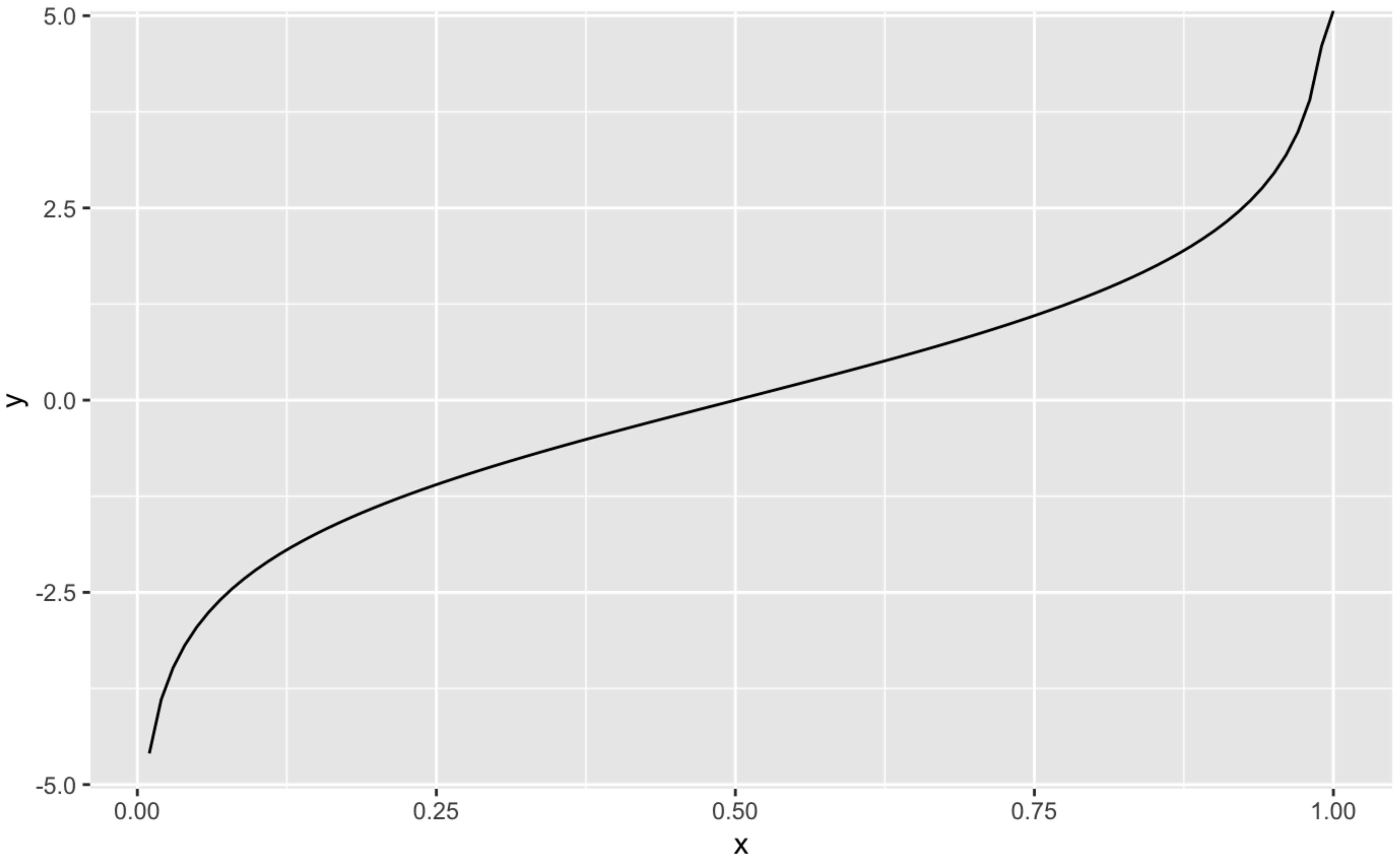
Logistic Regression: Odds?

- Running normal linear regression with a binary response variable can lead to nonsensical outcomes (e.g., negative probabilities)
- We need to come up with a different response variable
- Let $p = \Pr(D)$ be the probability of having diabetes
- Idea 1: Odds of having diabetes: $\dfrac{p}{1-p} \in [0,\infty]$



Logistic Regression: Log-Odds

- Idea 2: Log-odds of having diabetes: $\ln\left(\frac{p}{1-p}\right) \in [-\infty, \infty]$



Logistic Regression: Log-Odds

- Log-odds of having diabetes: $\ln\left(\frac{p}{1-p}\right) \in [-\infty, \infty]$
- Maps $[0,1]$ to $[-\infty, \infty]$
- Given a log-odds value of x, we can get $p = \Pr(D)$ back as follows:

$$p = \frac{e^x}{1 + e^x}$$

- With the log-odds function, we can run "normal" regression now

Logistic Regression: R

```
```{r}
logit1 <- glm(diabetes~weight, data=diabetes_data, family="binomial")
summary(logit1)
```
```

```
Call:
glm(formula = diabetes ~ weight, family = "binomial", data = diabetes_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.6557  -0.6705  -0.5614   0.8154   2.0965  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -6.75790    1.94453  -3.475  0.00051 ***
weight       0.03898    0.01190   3.275  0.00106 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65.342  on 49  degrees of freedom
Residual deviance: 51.925  on 48  degrees of freedom
AIC: 55.925

Number of Fisher Scoring iterations: 4
```

Logistic Regression: Interpretation

- Model:

$$\log\left(\frac{p}{1-p}\right) = -6.758 + 0.039 \cdot \text{weight}$$

- Log-odds of a 165 lb person having diabetes:

$$\log\left(\frac{p}{1-p}\right) = -6.758 + 0.039 \cdot 165 = -0.323$$

$$p = \frac{e^{-0.323}}{1 + e^{-0.323}} = 0.420$$

- Interpretation of slope: Change in log-odds ratio per unit change in predictor (in this case, weight) – often relatively unintuitive

Logistic Regression: Evaluation

- How do we evaluate if the model is effective?
  - Null hypothesis: Explanatory variables do not help explain log-odds response ("model is not useful")
  - Alternative hypothesis: At least one explanatory variable helps explain log-odds response ("model is useful")
- Deviance:
  - Null deviance (total variability around mean) minus residual deviance (error in prediction) follows a χ^2 distribution with p degrees of freedom
  - Test statistic: X^2 = null deviance - residual deviance (both given by R)
  - Calculate p-value: $p = \Pr(\chi_p^2 > X^2)$ – here, we care about the upper tail p-value
  - Conclusion: If $p < \alpha$, we reject the null hypothesis and conclude that the model is

Logistic Regression: Comparing Multiple Models

- We can compare the efficacy of multiple models through *information criterion* approaches (measure tradeoff of *explanatory power* and *simplicity*)
    - AIC: Akaike information criterion
    - BIC: Bayesian information criterion
- Must compare AIC or BIC values for different models on the *same* data (absolute values are not comparable between different settings)
    - Lower is better
    - Still must check residuals! Lowest AIC/BIC score doesn't mean the model is necessarily good (all models may be bad)

Logistic Regression: R

```{r}
logit1 <- glm(diabetes~weight, data=diabetes_data, family="binomial")
summary(logit1)
```

```
Call:
glm(formula = diabetes ~ weight, family = "binomial", data = diabetes_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6557  -0.6705  -0.5614   0.8154   2.0965

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.75790    1.94453  -3.475  0.00051 ***
weight       0.03898    0.01190   3.275  0.00106 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65.342  on 49  degrees of freedom
Residual deviance: 51.925  on 48  degrees of freedom
AIC: 55.925

Number of Fisher Scoring iterations: 4
```

```{r}
logit2 <- glm(diabetes~height, data=diabetes_data, family="binomial")
summary(logit2)
```

```
Call:
glm(formula = diabetes ~ height, family = "binomial", data = diabetes_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3582  -0.8968  -0.6762   1.1324   1.9529

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.85616    4.78450  -2.269   0.0233 *
height        0.15441    0.07136   2.164   0.0305 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65.342  on 49  degrees of freedom
Residual deviance: 60.039  on 48  degrees of freedom
AIC: 64.039

Number of Fisher Scoring iterations: 4
```

Logistic Regression: Inference for Coefficients

- We can also run hypothesis tests for each coefficient independently
- Hypotheses: $H_0 : \hat{\beta}_j = \beta_j^*$ vs. $H_1 : \hat{\beta}_j \neq \beta_j^*$
- Evaluate using a *z-score* (normal distribution, not a t distribution)
    - Don't have to use sample errors to estimate variance; binomial distribution only has one parameter that underlies both mean and variance
- $z = \dfrac{\hat{\beta}_j - \beta_j^*}{SE(\hat{\beta}_j)}$ (standard error is given in the logistic regression output)
- $p = \Pr(|Z| > |z|) =$ `2*pnorm(-abs(z))`
- If $p < \alpha$, reject H_0

Logistic Regression: Interactions

```{r}
logit2 <- glm(diabetes~weight+height+sex+weight*sex, data=diabetes_data, family="binomial")
summary(logit2)
```

```
Call:
glm(formula = diabetes ~ weight + height + sex + weight * sex,
    family = "binomial", data = diabetes_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5068  -0.6900  -0.4197   0.8994   2.5298

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -14.95394   10.19871  -1.466    0.143
weight        0.10436    0.06891   1.514    0.130
height       -0.01234    0.09401  -0.131    0.896
sexM         14.77105   11.11219   1.329    0.184
weight:sexM  -0.09556    0.07487  -1.276    0.202

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 65.342  on 49  degrees of freedom
Residual deviance: 49.841  on 45  degrees of freedom
AIC: 59.841

Number of Fisher Scoring iterations: 5
```

Generalized Linear Models

- Logistic regression is an example of what is called a generalized linear model
- Generalized linear models:
  - Probability distribution describing the outcome variable
    - For logistic regression: binomial with parameter p
  - A linear model $\eta = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$
  - A function g that relates the linear model to the parameter of the outcome distribution
    - $g(p) = \eta$ or $p = g^{-1}(\eta)$
    - For logistic regression: $g(p) = logit(p) = \log\left(\frac{p}{1-p}\right)$
- In R: `glm()` command