Normality inducing transformations

Adding missing feature as a normality inducing transformation

The Linear Regression model is

$$\mathbf{y} = \Theta^T \mathbf{x} + \epsilon$$

As explained before, Regression produces a conditional probability

$$p(\hat{\mathbf{y}}|\mathbf{x})$$

where $\hat{\mathbf{y}}$ and ϵ are *Normally distributed variables.*

Assumptions of the Linear Regression model are violated if

- ϵ is not Normal
- the individual $\epsilon^{(\mathbf{i})}$ display a pattern
- the individual $\epsilon^{(\mathbf{i})}$ have different variances (heteroscedastic)

One reason for failure of these assumptions is a missing feature

- "curvy" data set and Linear model
    - we saw pattern of errors: larger in tails
    - variances increased in tail

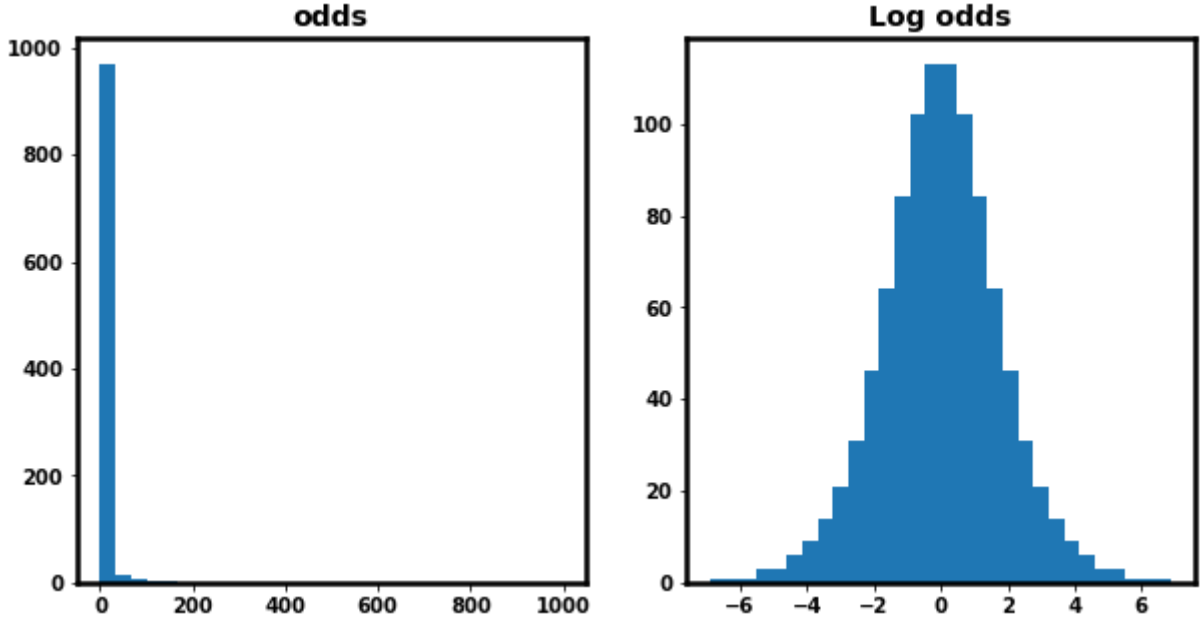
Adding a feature (e.g.,second order polynomial term for the curvy data set) can be seen as a normality inducing transformation.

Log transformation

We've seen this in our lecture on Logistic Regression

- the probabilities are *not* normally distributed
- the odds are *not* normally distributed
- the *log odds* **is** normally distributed

In [4]:
```
tf = tmh.TransformHelper()
tf.plot_odds()
```



$$
\begin{aligned}
\frac{\hat{p}}{1-\hat{p}} &= \frac{\frac{1}{1+e^{-s}}}{1-\frac{1}{1+e^{-s}}} \\
&= \frac{\frac{1}{1+e^{-s}}}{\frac{1+e^{-s}-1}{1+e^{-s}}} \\
&= \frac{1}{e^{-s}} \\
&= e^{s}
\end{aligned}
$$

So `LogisticRegression` is really just a `LinearRegression` with a transformed target

$$\log(\frac{\hat{p}}{1-\hat{p}}) = \Theta^T \cdot x$$

Other transformations

Centering

Transforming a feature to have mean 0.

$$\mathbf{x}_j^{(\mathbf{i})} = \mathbf{x}_j^{(\mathbf{i})} - \bar{\mathbf{x}}_j$$

- low values now become negative
    - more clearly indicates deleterious effect than a low, positive number
    - example: Star Ratings for movies
- some algorithms (PCA) need centered data

Bucketing/Binning

- Target may be linear in a feature only in broad ranges of the feature
    - income vs age
        - very young (below working age) all income is identical (0)
        - very old (above retirement) - no job related income
    - Latitude/Longitude
        - small changes matter MUCH less than big changes
- Converts numerical feature
    - into categorical $\mathbf{Is\ bucket\ 1}, \mathbf{Is\ bucket\ 2}, \ldots$
    - ordinal: replace value with center value of bin

Bucket size choices:

- Equal spaced buckets
- Equal quantile buckets

Lesson Don't fit a square peg (non-linear response) into a round hole (linear model)

Outliers

Pull in extreme values to reduce their influence on the fit.

- Clipping, Winsorization

In [9]:
```
print("Done")
```

```
Done
```