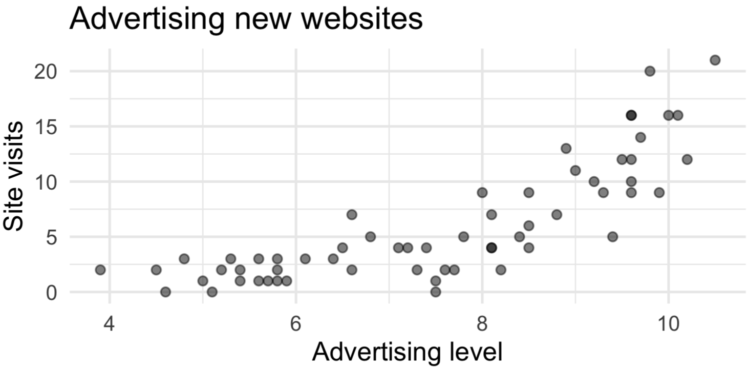
Andrew Gard - equitable.equations@gmail.com

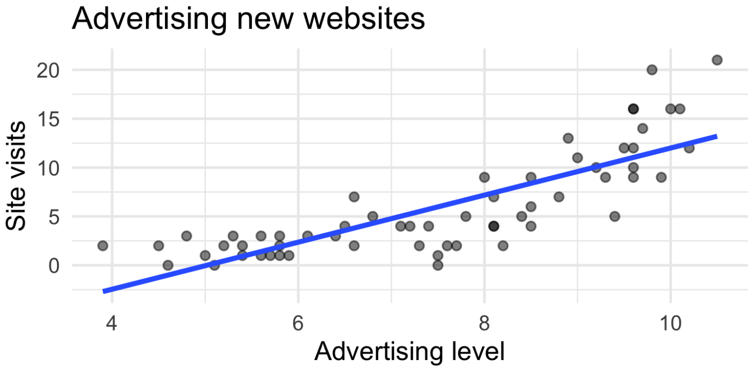
Poisson regression

Poisson regression is used to model *counts*, for instance the number of eggs laid by fish or the number of customers at a small business. Explanatory variables can be quantitative or categorical.

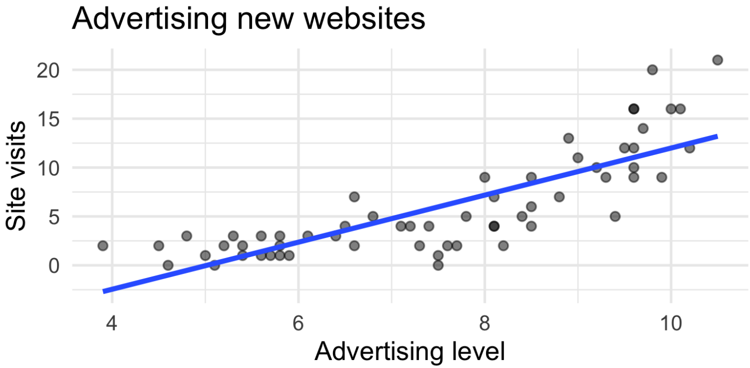
Poisson regression is used to model *counts*, for instance the number of eggs laid by fish or the number of customers at a small business. Explanatory variables can be quantitative or categorical. For instance:



Linear regression is the wrong tool in problems like this.

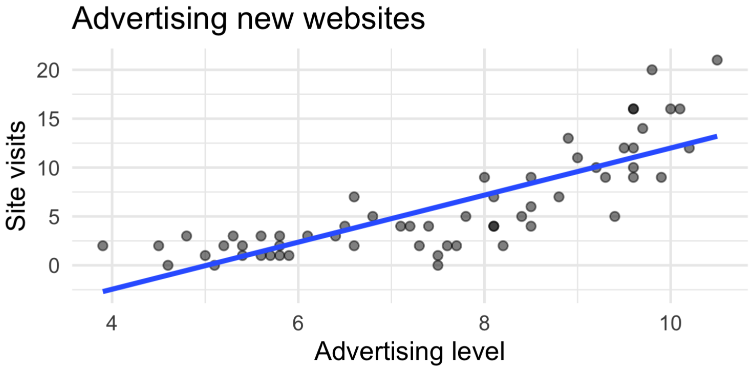


Linear regression is the wrong tool in problems like this.



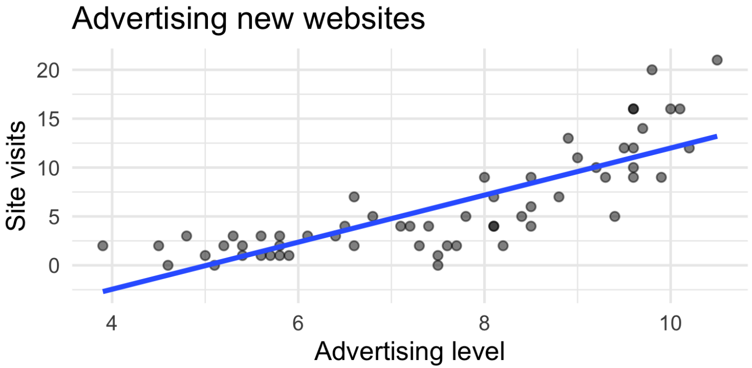
- The shape of the plot isn't linear.

Linear regression is the wrong tool in problems like this.



- The shape of the plot isn't linear.
- A regression line may predict negative counts for some x-values.

Linear regression is the wrong tool in problems like this.



- The shape of the plot isn't linear.
- A regression line may predict negative counts for some x-values.
- The data becomes more variable as the x-value increases.

The poisson model is designed to fit this sort of data. In its simplest form (one quantitative explanatory variable), it looks like this:

$$\begin{cases} \ln(\lambda) &= \beta_0 + \beta_1 x & \text{systematic component} \\ y &\sim \text{pois}(\lambda) & \text{random component} \end{cases}$$

where λ is the expected number of occurrences when the value of the explanatory variable is x.

The poisson model is designed to fit this sort of data. In its simplest form (one quantitative explanatory variable), it looks like this:

$$\begin{cases} \ln(\lambda) = \beta_0 + \beta_1 x & \text{systematic component} \\ y \sim \text{pois}(\lambda) & \text{random component} \end{cases}$$

where λ is the expected number of occurrences when the value of the explanatory variable is x. β_0 and β_1 are unknown coefficients that must be estimated using data.

The poisson model is designed to fit this sort of data. In its simplest form (one quantitative explanatory variable), it looks like this:

$$\begin{cases} \ln(\lambda) &= \beta_0 + \beta_1 x & \text{systematic component} \\ y &\sim \text{pois}(\lambda) & \text{random component} \end{cases}$$

where λ is the expected number of occurrences when the value of the explanatory variable is x. β_0 and β_1 are unknown coefficients that must be estimated using data.

Because the logarithm function is invertible, the systematic component can also be written:

$$\lambda = e^{\beta_0 + \beta_1 x}$$

The model is fit using the method of maximum likelihood estimation.

The model is fit using the method of maximum likelihood estimation. The math is ugly and is *always* done with technology. For instance, in R, the code to obtain the coefficients for the advertising data in the previous slides is

$$\texttt{glm}(\texttt{count} \sim \texttt{level}, \texttt{data} = \texttt{advertising}, \texttt{family} = \text{``}\texttt{poisson}\text{''})$$

The model is fit using the method of maximum likelihood estimation. The math is ugly and is *always* done with technology. For instance, in R, the code to obtain the coefficients for the advertising data in the previous slides is

$$\texttt{glm}(\texttt{count} \sim \texttt{level}, \texttt{data} = \texttt{advertising}, \texttt{family} = \text{“}\texttt{poisson}\text{”})$$

The resulting printout is:

```
Coefficients:
(Intercept)        level
    -2.1227       0.4778

Degrees of Freedom: 59 Total (i.e. Null);  58 Residual
Null Deviance:      264.8
Residual Deviance: 64.84        AIC: 261.8
```

In this example, the maximum likelihood estimators for β_0 and β_1 are

$$\hat{\beta}_0 = -2.1227 \quad \text{and} \quad \hat{\beta}_1 = .4778$$

In this example, the maximum likelihood estimators for β_0 and β_1 are

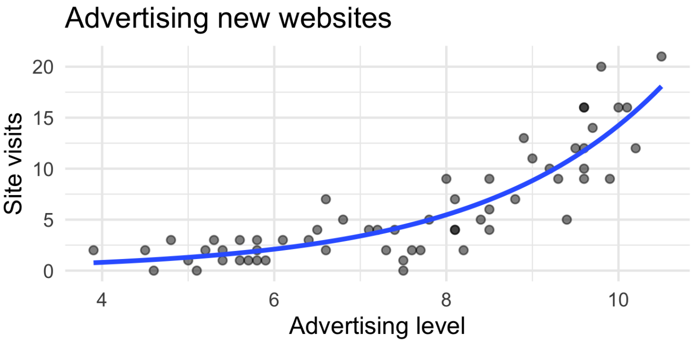
$$\hat{\beta}_0 = -2.1227 \quad \text{and} \quad \hat{\beta}_1 = .4778$$

This gives us the following model,

$$\begin{cases} \ln(\hat{\lambda}) & = & -2.12 + .48x & \text{systematic component} \\ y & \sim & \text{pois}(\hat{\lambda}) & \text{random component} \end{cases}$$

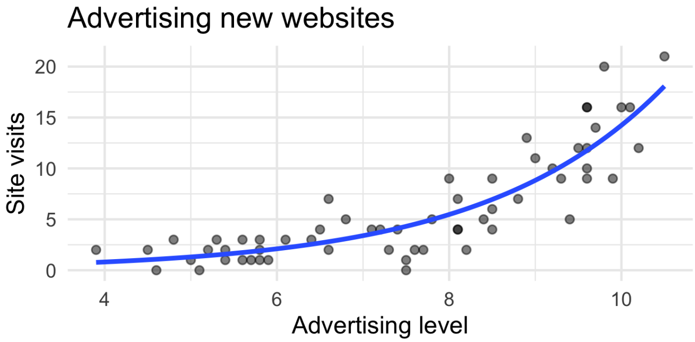
This model fits the data very well.

$$\hat{\lambda} = e^{-2.12+.48x}$$



This model fits the data very well.

$$\hat{\lambda} = e^{-2.12+.48x}$$



For instance, the model says that at an advertising level of $x = 8$, the expected number of site visits is $\hat{\lambda} = e^{-2.12+.48 \cdot 8} \approx 5.6$. Individual counts at that level will have a poisson distribution with this

The coefficients in a poisson regression model are only slightly more difficult to interpret than the coefficients in a linear one.

$$\lambda = e^{\beta_0 + \beta_1 x}$$

The coefficients in a poisson regression model are only slightly more difficult to interpret than the coefficients in a linear one.

$$\begin{aligned} \lambda &= e^{\beta_0 + \beta_1 x} \\ &= e^{\beta_0} \cdot (e^{\beta_1})^x \end{aligned}$$

The coefficients in a poisson regression model are only slightly more difficult to interpret than the coefficients in a linear one.

$$
\begin{aligned}
\lambda &= e^{\beta_0+\beta_1 x} \\
&= e^{\beta_0} \cdot (e^{\beta_1})^x
\end{aligned}
$$

- The predicted mean count when $x = 0$ is e^{β_0}.

The coefficients in a poisson regression model are only slightly more difficult to interpret than the coefficients in a linear one.

$$\begin{aligned}\lambda &= e^{\beta_0+\beta_1 x} \\ &= e^{\beta_0} \cdot (e^{\beta_1})^x\end{aligned}$$

- The predicted mean count when $x = 0$ is e^{β_0}.
- An increase of one unit in the explanatory variable increases the predicted mean count by a *factor* of e^{β_1}.

The coefficients in a poisson regression model are only slightly more difficult to interpret than the coefficients in a linear one.

$$\begin{aligned} \lambda &= e^{\beta_0+\beta_1 x} \\ &= e^{\beta_0} \cdot (e^{\beta_1})^x \end{aligned}$$

- The predicted mean count when $x = 0$ is e^{β_0}.
- An increase of one unit in the explanatory variable increases the predicted mean count by a *factor* of e^{β_1}.

In the advertising example, the predicted mean count when $x = 0$ is $e^{-2.12} \approx 0.12$ and an increase of one unit in x increases the predicted mean count by a factor of $e^{.48} \approx 1.62$.

There are a number of important differences between poisson regression and linear regression.

There are a number of important differences between poisson regression and linear regression.

- There is no normal distribution in a poisson regression model. At every x-value, the distribution of the response variable is *poisson*.

There are a number of important differences between poisson regression and linear regression.

- There is no normal distribution in a poisson regression model. At every x-value, the distribution of the response variable is *poisson*.
- The variance of the response variable isn't constant in a poisson regression model. In fact, it's larger for larger values of λ.

There are a number of important differences between poisson regression and linear regression.

- There is no normal distribution in a poisson regression model. At every x-value, the distribution of the response variable is *poisson*.
- The variance of the response variable isn't constant in a poisson regression model. In fact, it's larger for larger values of λ.
- Raw residuals (the differences between the observed and expected y-values) aren't a good measure of model fit since larger errors are expected for larger λ. Deviance and Pearson (standardized) residuals are typically used instead.