Module 7: Generalized Linear Models

ISE-529

Material largely drawn from ISLR Chapter 4.6

Generalized Linear Models

Overview

- Thus far, we have considered two types of response variables:
  - Quantitative (measures)
  - Qualitative (categories)
- Not all types of models fit neatly into these two categories

Generalized Linear Models

Overview

We will look at the Bikeshare data set

- Response variable: “bikers” – number of hourly users of a bike sharing program in Washington, DC
- Predictors:
  - month (month of the year)
  - hr (hour of the day from 0-23)
  - temp (normalized temperature in Celsius)
  - weathersit (qualitative variable with one of four possible values – “clear”, “misty or cloudy”, “light rain or light snow”, “heavy rain or heavy snow”

Generalized Linear Models

Bikeshare Dataset

Predictor variables

Response variable

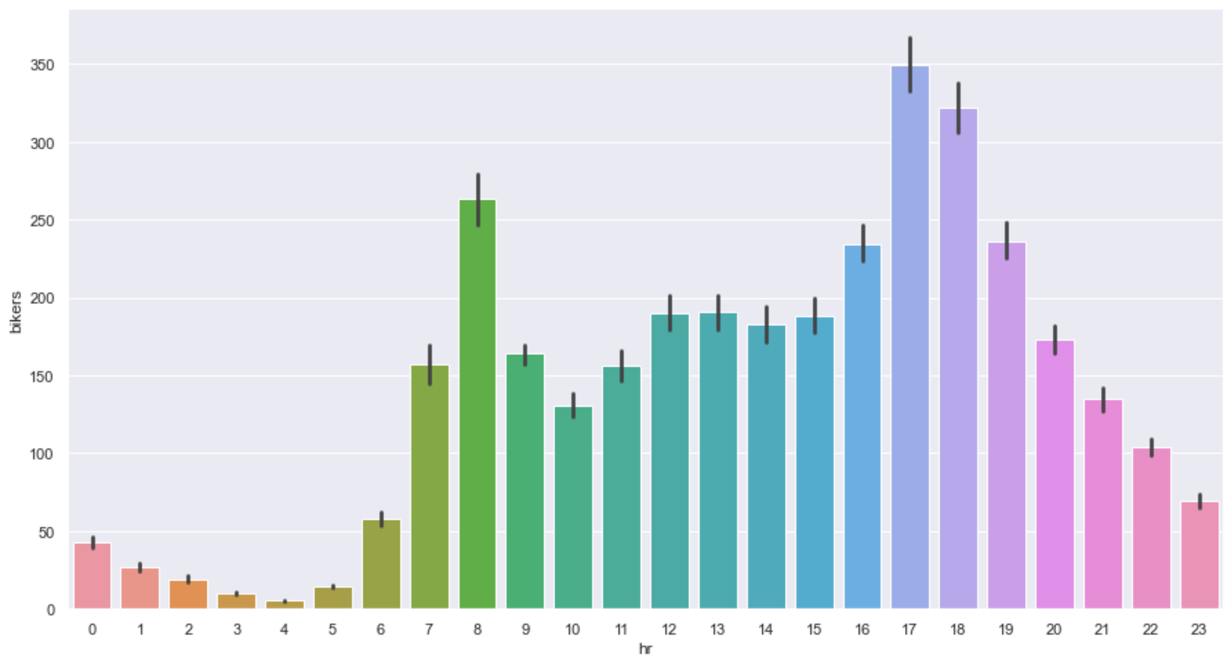
	season	mnth	day	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	bikers
0	1	Jan	1	0	0	6	0	clear	0.24	0.2879	0.81	0.0000	3	13	16
1	1	Jan	1	1	0	6	0	clear	0.22	0.2727	0.80	0.0000	8	32	40
2	1	Jan	1	2	0	6	0	clear	0.22	0.2727	0.80	0.0000	5	27	32
3	1	Jan	1	3	0	6	0	clear	0.24	0.2879	0.75	0.0000	3	10	13
4	1	Jan	1	4	0	6	0	clear	0.24	0.2879	0.75	0.0000	0	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
8640	1	Dec	365	19	0	6	0	clear	0.42	0.4242	0.54	0.2239	19	73	92
8641	1	Dec	365	20	0	6	0	clear	0.42	0.4242	0.54	0.2239	8	63	71
8642	1	Dec	365	21	0	6	0	clear	0.40	0.4091	0.58	0.1940	2	50	52
8643	1	Dec	365	22	0	6	0	clear	0.38	0.3939	0.62	0.1343	2	36	38
8644	1	Dec	365	23	0	6	0	clear	0.36	0.3788	0.66	0.0000	4	27	31

8645 rows × 15 columns

Bikeshare Dataset

Set Up Linear Regression Model

Would you treat “hour” as a category or a measure?



Bikeshare Dataset

Linear Regression Model

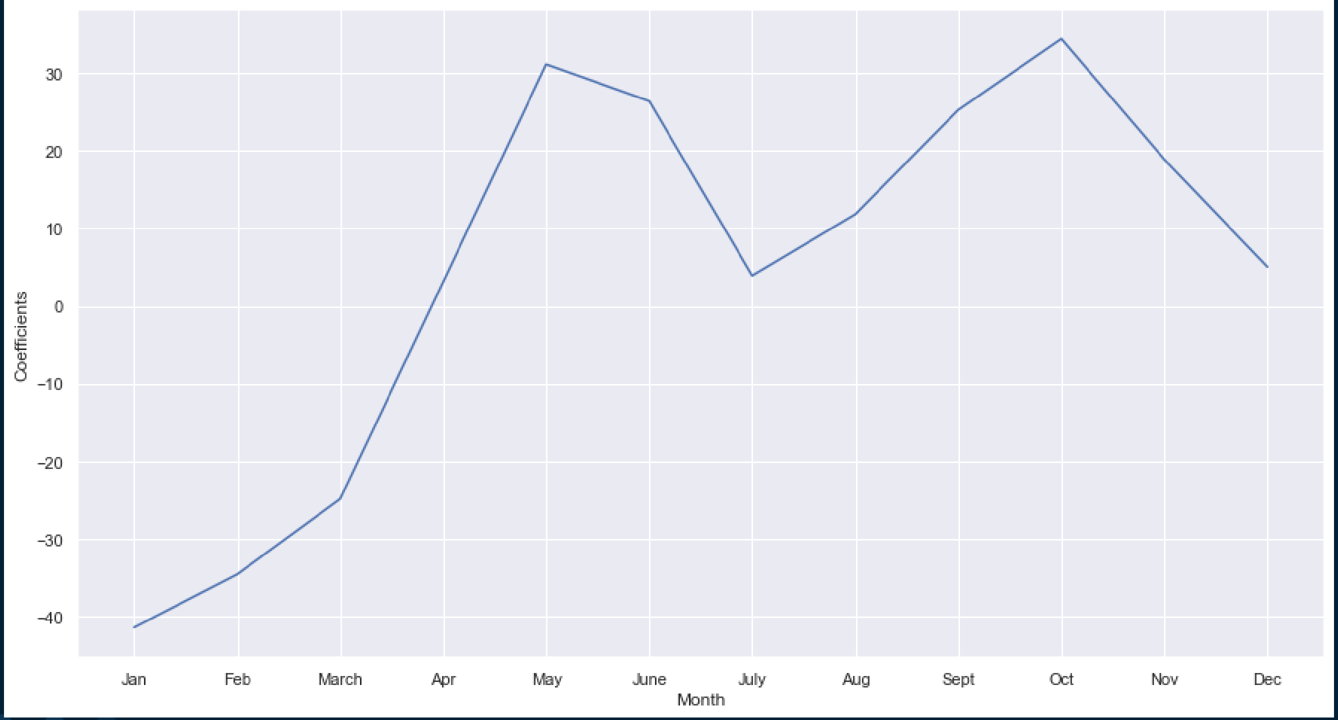
```python
lrmodel1 = LinearRegression(fit_intercept = True)
lrmodel1.fit(X,y)
lrmodel1_coefs = pd.DataFrame(lrmodel1.coef_, columns = ['Coefficients'], index = X.columns)
lrmodel1_coefs.loc[['workingday', 'temp', 'weathersit_cloudy/misty', 'weathersit_heavy rain/snow',
                    'weathersit_light rain/snow']]
```

	Coefficients
workingday	1.269601
temp	157.209366
weathersit_cloudy/misty	-12.890266
weathersit_heavy rain/snow	-109.744577
weathersit_light rain/snow	-66.494365

Does this look reasonable?

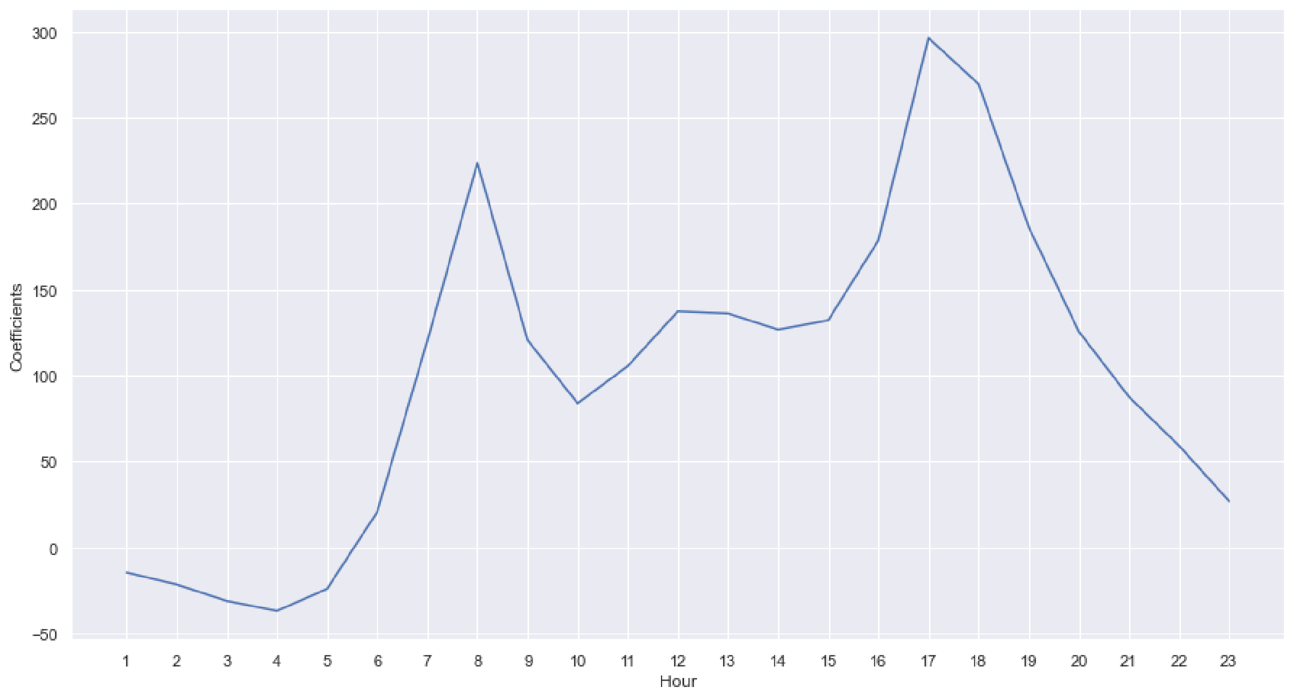
Bikeshare Dataset

Linear Regression Model – Month Coefficients



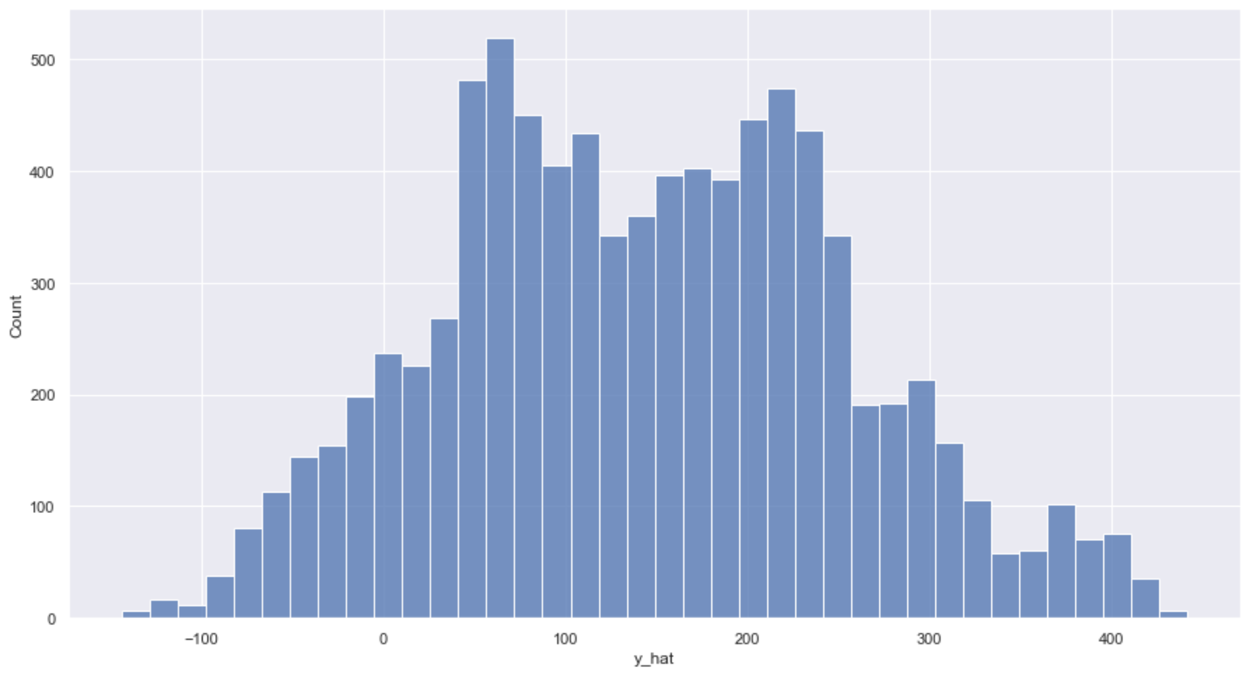
Bikeshare Dataset

Linear Regression Model – Hour Coefficients



Bikeshare Data

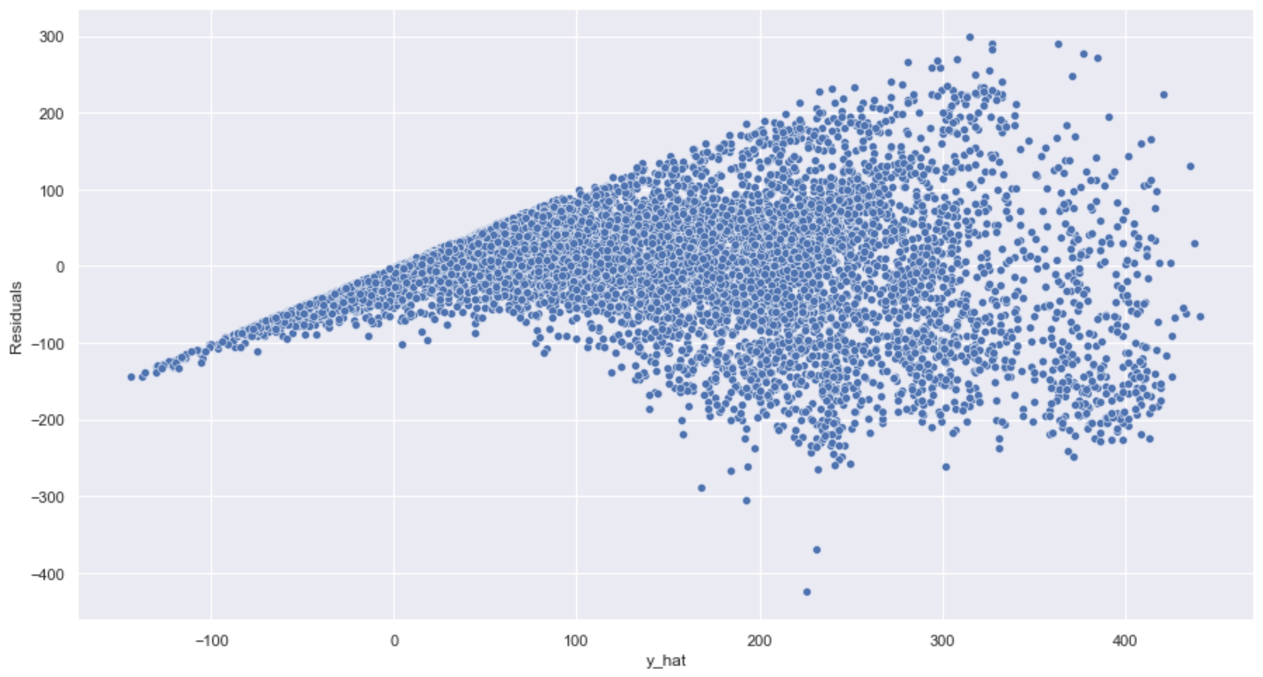
Linear Regression Model - Predictions Histogram



What issue do you see?

Bikeshare Data

Linear Regression Model - Residuals

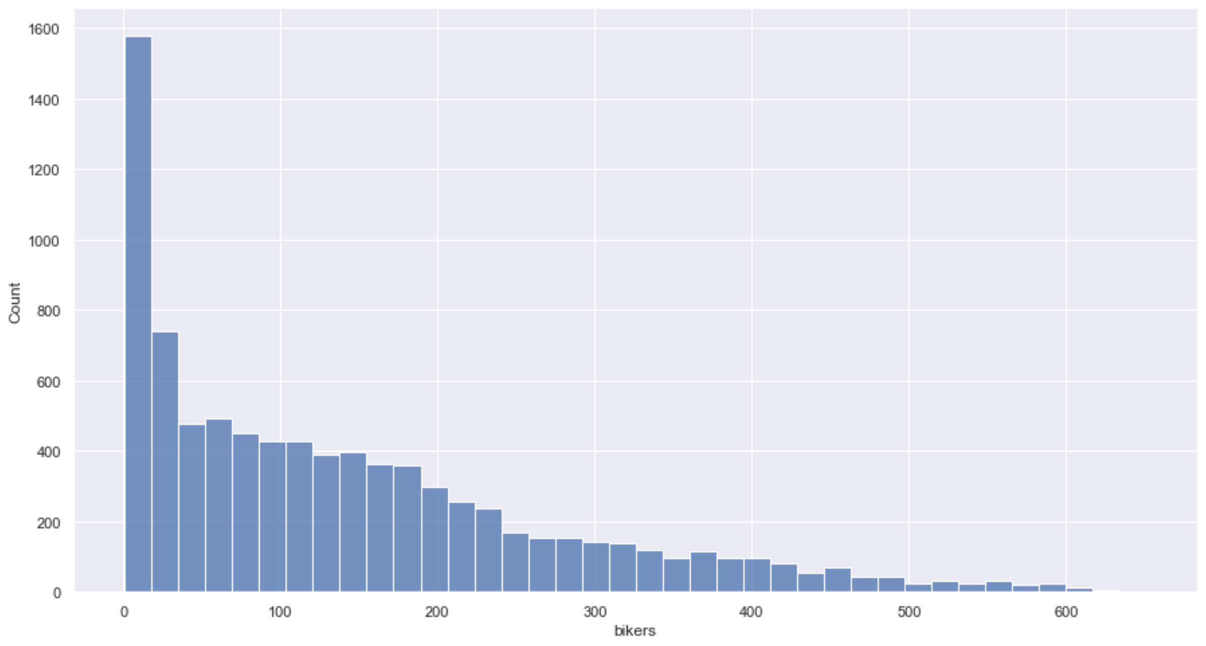


What issue do you see?

Bikeshare Data

Try Transforming the Response Variable

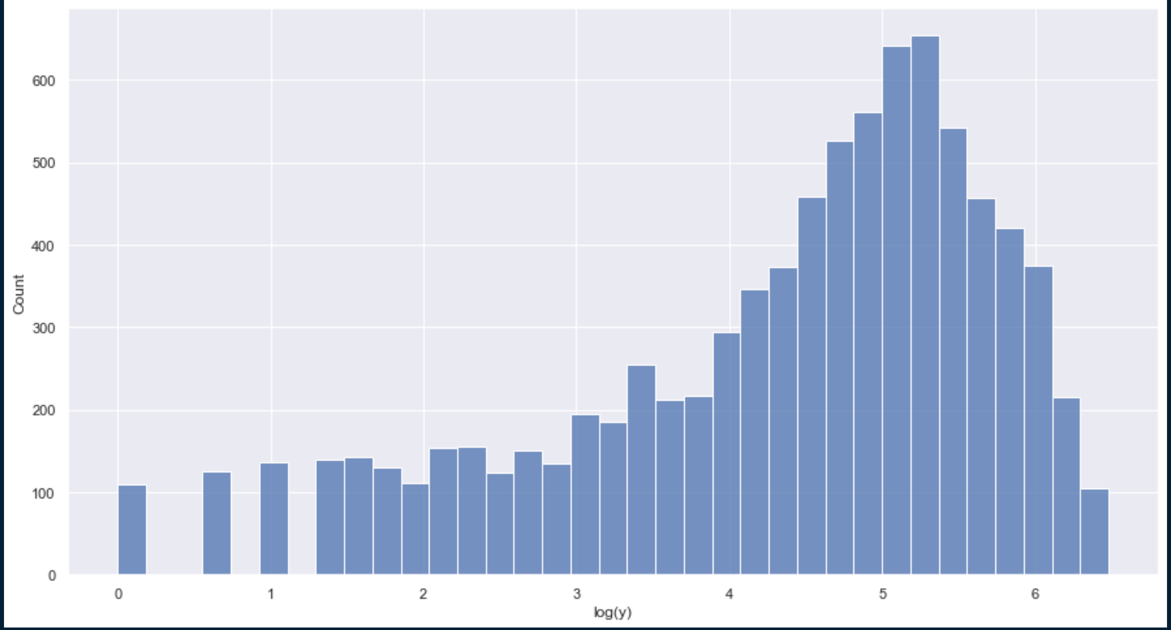
Looking at the response variable, we find that it is significantly skewed:



Bikeshare Data

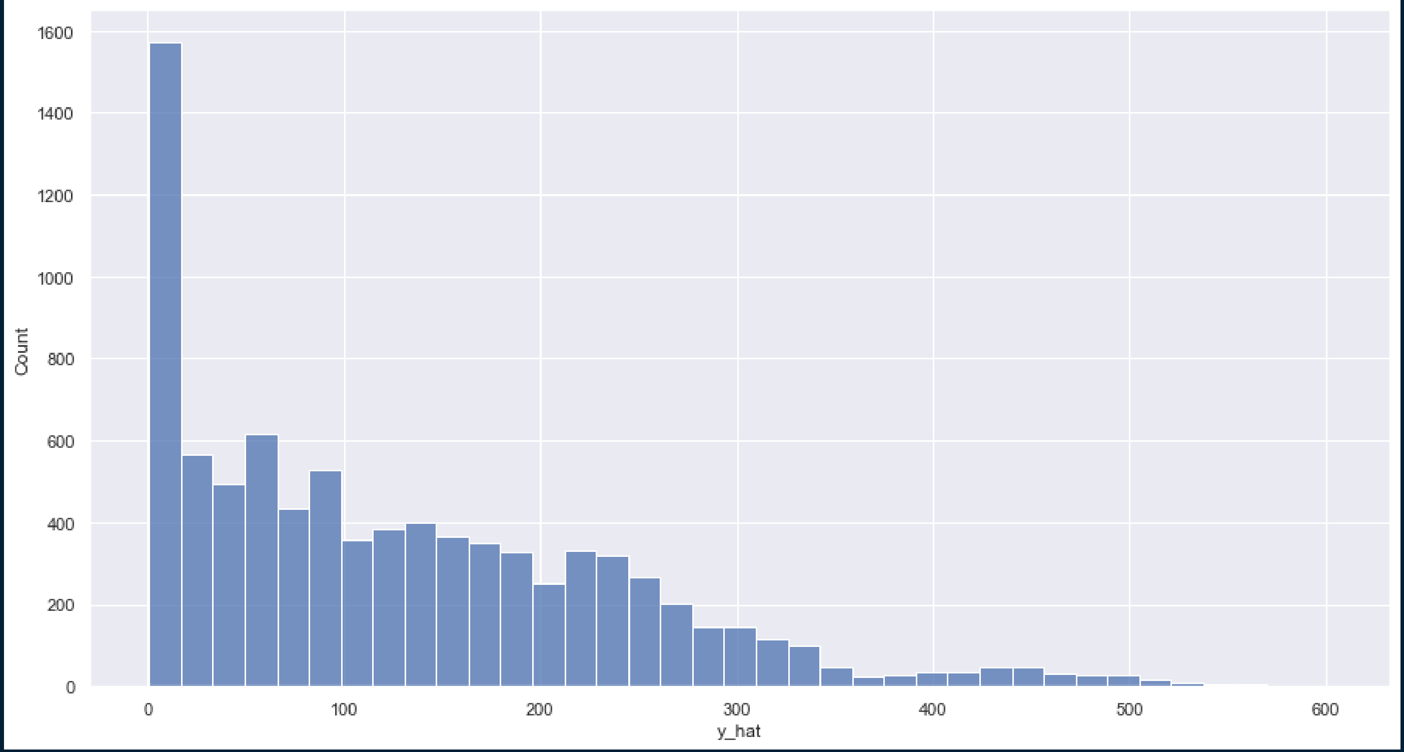
Try Transforming the Response Variable

A log transform significantly improves the skew:



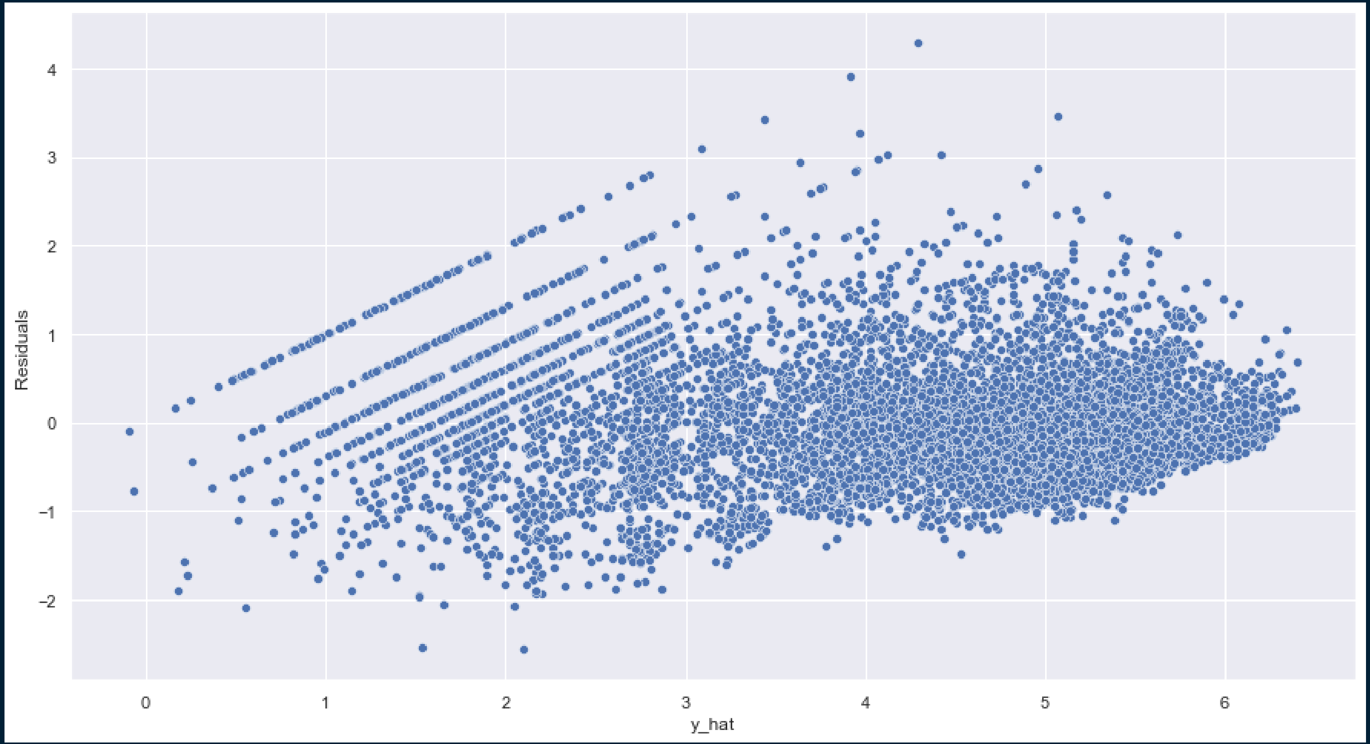
Bikeshare Data

Linear Regression Model With Transformed Response



Bikeshare Data

Linear Regression Model With Transformed Response



Poisson Regressions

A Better Approach

Reminders: Poisson distribution

- A discrete, non-negative distribution that is often used to model counts

$$P(Y=k)=\frac{e^{-\lambda}\lambda^{k}}{k!} \qquad \text{for } k=0,1,2,\ldots$$

Poisson Regression

Reminders, for Poisson distributions:

- $E(Y) = \lambda$

If we model bikers with a Poisson distribution with $E(Y) = \lambda = 5$, then for a particular hour:

- $P(Y = 0) = \frac{e^{-5}5^0}{0!} = e^{-5}$ = 0.0067
- $P(Y = 1) = \frac{e^{-5}1}{1!} = 5e^{-5}$ = 0.034
- $P(Y = 2) = \frac{e^{-5}2}{2!}$ = 0.084
- …

However, we expect the mean number of users in an hour to vary as a function of the hour of the day, month of the year, weather conditions, etc.

Poisson Regression

Basic Approch

Rather than modeling the number of bikers as a Poisson distribution with a fixed mean value (such as $\lambda = 5$), we model the mean as a function of the predictor variables:

$$\log(\lambda(X_1, \ldots, X_p) = \beta_0 + \beta_1 X_1, \ldots, \beta_p X_p$$

or, equivalently:

$$\lambda(X_1, \ldots, X_p) = \chi^{\beta_0 + \beta_1 X_1, \ldots, \beta_p X_p}$$

Poisson Regression

Modeling the Bikeshare Dataset

Poisson regression

```
from sklearn.linear_model import PoissonRegressor
prmodel_1 = PoissonRegressor()
prmodel_1.fit(X,y)
prmodel1_coefs_df = pd.DataFrame(prmodel_1.coef_, columns = ['Coefficients'], index = X.columns)
prmodel1_coefs_df.loc[['workingday', 'temp', 'weathersit_cloudy/misty', 'weathersit_heavy rain/snow',
                    'weathersit_light rain/snow']]
```

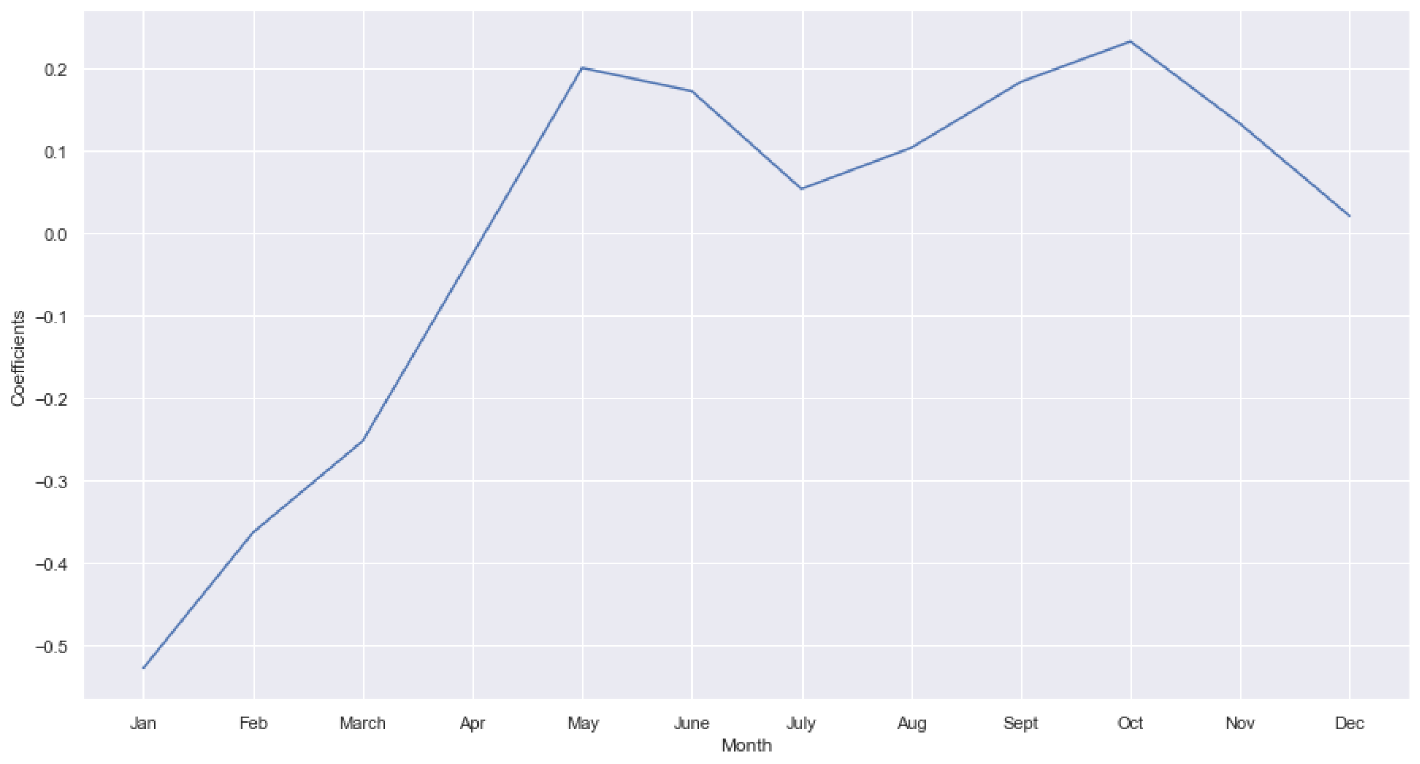
	Coefficients
workingday	0.011061
temp	0.919983
weathersit_cloudy/misty	-0.063318
weathersit_heavy rain/snow	-0.006610
weathersit_light rain/snow	-0.491935

```
prmodel_1.intercept_
```

4.191501499376311

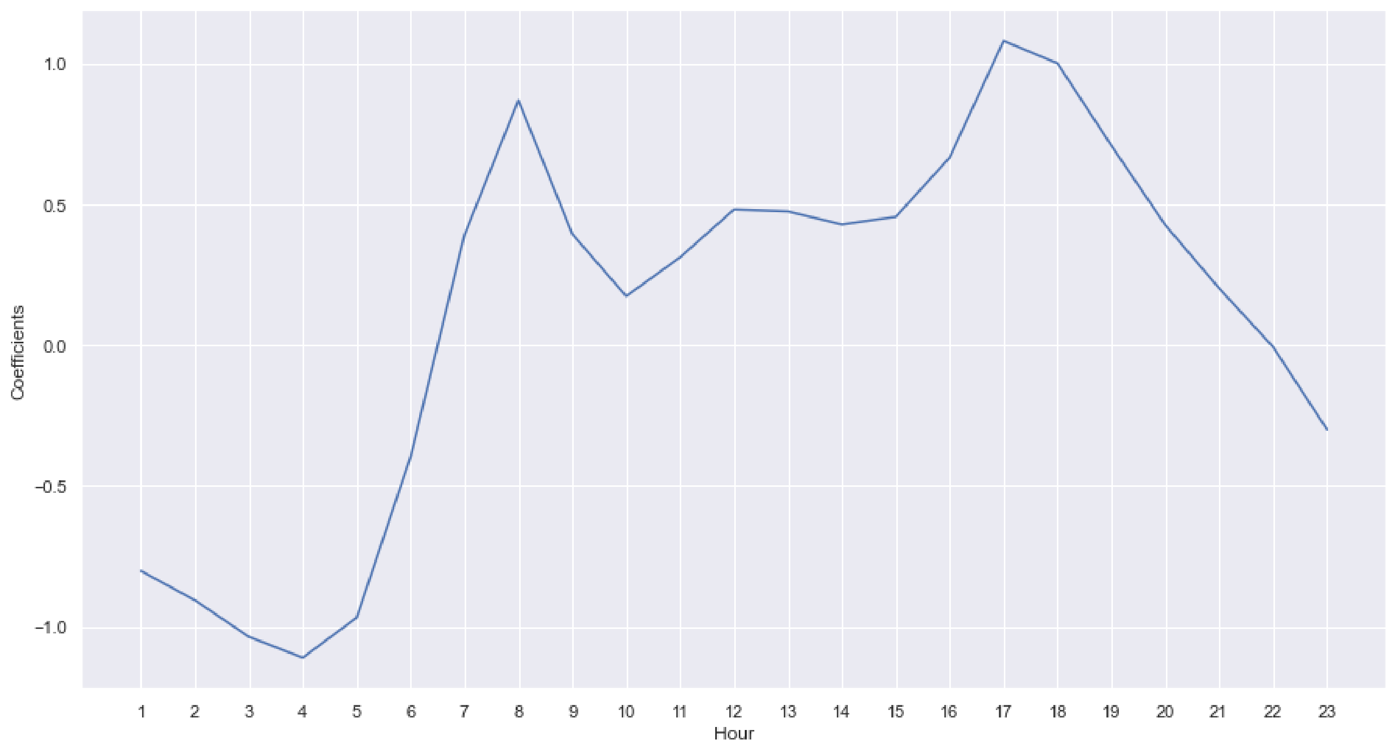
Poisson Regression

Modeling the Bikeshare Dataset



Poisson Regression

Modeling the Bikeshare Dataset



Poisson Regression

Inference

Reminder: $\lambda(X_1, \dots, X_p) = e^{\beta_0+\beta_1 X_1+,\dots,+\beta_p X_p}$

- Thus, an increase by 1 in any predictor value j causes an increase in the mean value of the response variable by a factor of e^j (holding all other predictors constant)
  - A change from clear to cloudy weather changes the mean bike usage by a factor of $e^{-.06}$ = 0.94. On average only 94% as many people use bikes on a cloudy day compared to when it is clear
  - A change from clear to light rain changes the mean by a factor of $e^{-0.5}$ = .607 (only 60% as many use bikes on a rainy day than a clear day)

Poisson Regression

Mean-Variance Relationship

- In a Poisson distribution, $Var(Y) = \lambda$
- By using this distribution, the assumption is that variance increases as the mean increases
  - Different from linear regression models where the assumption is that the variance is constant and independent of the mean
- Thus, the Poisson regression is able to handle the mean-variance relationship generally seen in count variables in a natural way

Generalized Linear Models

Linear Model Types

We have now looked at three types of linear models

- Linear regression
- Logistic regression
- Poisson regression

Linear Model Types

Commonality of the three approaches:

- Each uses predictors $X_1, \ldots, X_p$ to predict a response Y
- We assume that conditional on $X_1, \ldots, X_p$, Y belongs to a certain family of distributions:

Model	Distribution Family for Y
Linear regression	
Logistic regression	
Poisson regression	

Linear Model Types

Commonality of the three approaches:

- Each uses predictors $X_1, \ldots, X_p$ to predict a response Y
- We assume that conditional on $X_1, \ldots, X_p$, Y belongs to a certain family of distributions:

Model	Distribution Family for Y
Linear regression	Gaussian (normal)
Logistic regression	Bernoulli
Poisson regression	Poisson

Linear Model Types

Commonality of the three approaches:

- Each models the mean of Y as a function of the predictors

Model	Distribution Family for Y
Linear regression	$\beta_0 + \beta_1 X_1 +, \dots, +\beta_p X_p$
Logistic regression	$\frac{e^{\beta_0 + \beta_1 X_1 +, \dots, +\beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 +, \dots, +\beta_p X_p}}$
Poisson regression	$e^{\beta_0 + \beta_1 X_1 +, \dots, +\beta_p X_p}$

Linear Model Types

Link Functions

Each of these three equations can be expressed using a “link function”:

$$\eta\left(E(Y|X_1, \ldots, X_p\right) = \beta_0 + \beta_1 X_1 +, \ldots, +\beta_p X_p$$

Thus, we have these link functions:

Model	Link function
Linear regression	$\eta(\mu) = \mu$
Logistic regression	$\eta(\mu) = \log(\frac{\mu}{1-\mu})$
Poisson regression	$\eta(\mu) = \log(\mu)$

Generalized Linear Model

Generalizes the linear regression model with two options:

- Link function
- Probability distribution of Y

Response Variable	Distribution	Link Function	Variance Function
Continuous	Normal	Identity	σ^2
Binary	Binomial	Logit	$\mu(1-\mu)$
Count	Poisson	Log	λ

Generalized Linear Model

Other Types of GLMs

- Gaussian, Bernoulli and Poisson distributions are all members of a class of distributions known as *exponential distributions*
- Other well-known exponential distributions are the exponential distribution, Gamma distribution, and negative binomial distribution
- GLMs model the response Y as coming from a particular member of the exponential family and then transforming the mean of the response so that the transformed mean is a linear function of the predictors
- Other examples of GLMs are *Gamma regression* and *negative binomial regressison*

Generalized Linear Models

Supported Linear Distributions (SAS)

Distribution	Available Link Functions (default listed first)
Beta	Logit, Probit, Log-log, C-log-log
Binary	Logit, Probit, Log-log, C-log-log
Exponential	Log, Identity
Gamma	Log, Identity, Reciprocal
Geometric	Log, Identity
Inverse Gaussian	Power(-2), Log, Identity
Negative Binomial	Log, Identity
Normal (default)	Identity, Log
Poisson	Log, Identity
Tweedie	Identity, Log