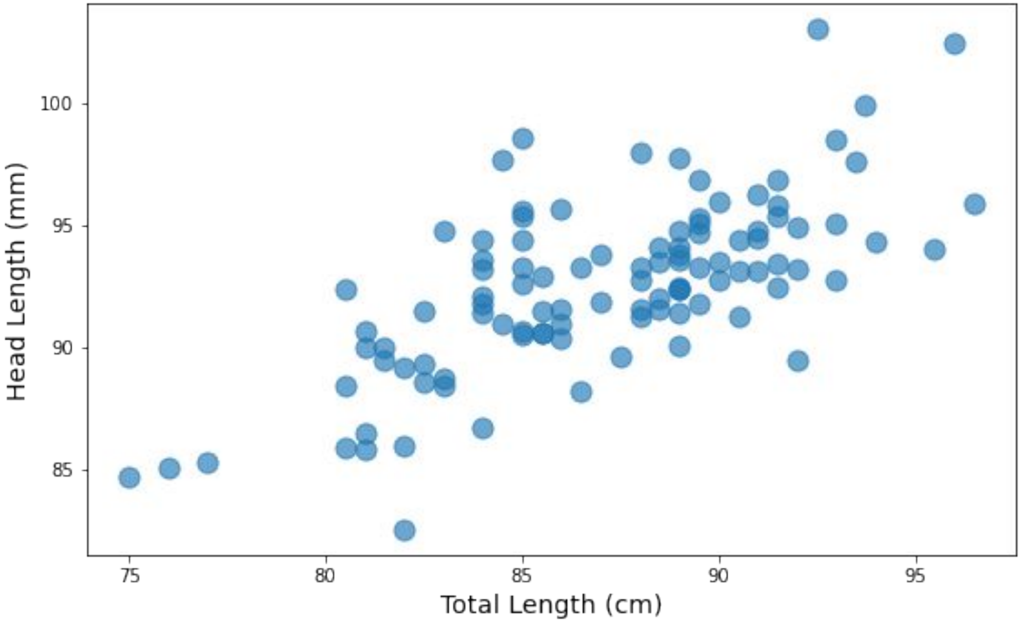
Introduction to Generalized Linear Models

Part 1: Linear Regression



Goal: Predict an Australian brushtail possum's head length

OpenIntro Statistics, Section 8.1.2

Approach 1: Ignore the Total Length variable, and just look at the overall distribution of Head Length.

```
1  possum.head()
```

	site	pop	sex	age	head_l	kull_w	total_l	tail_l
0	1	Vic	m	8.0	94.1	60.4	89.0	36.0
1	1	Vic	f	6.0	92.5	57.6	91.5	36.5
2	1	Vic	f	6.0	94.0	60.0	95.5	39.0
3	1	Vic	f	6.0	93.2	57.1	92.0	38.0
4	1	Vic	f	2.0	91.5	56.3	85.5	36.0

target column

Approach 1: Ignore the Total Length variable, and just look at the overall distribution of Head Length.

Approach 1: Ignore the Total Length variable, and just look at the overall distribution of Head Length.

```
import statsmodels.api as sm

linreg = (sm.GLM(endog = possum['head_l'],
                exog = sm.add_constant(possum[[]]),
                family = sm.families.Gaussian())
          .fit()
         )
```

Approach 1: Ignore the Total Length variable, and just look at the overall distribution of Head Length.

```
import statsmodels.api as sm

linreg = (sm.GLM(endog = possum['head_l'],
                exog = sm.add_constant(possum[[]]),
                family = sm.families.Gaussian())
          .fit()
         )
```

We'll be using the *statsmodels* library.

Approach 1: Ignore the Total Length variable, and just look at the overall distribution of Head Length.

```
import statsmodels.api as sm

linreg = (sm.GLM(endog = possum['head_l'],
                exog = sm.add_constant(possum[[]]),
                family = sm.families.Gaussian())
          .fit()
         )
```

Fit a Generalized Linear Model (GLM).

Approach 1: Ignore the Total Length variable, and just look at the overall distribution of Head Length.

```
import statsmodels.api as sm

linreg = (sm.GLM( endog = possum['head_l'] ,
                  exog = sm.add_constant(possum[[]]),
                  family = sm.families.Gaussian())
            .fit()
          )
```

This tells the model the target variable.

Approach 1: Ignore the Total Length variable, and just look at the overall distribution of Head Length.

```
import statsmodels.api as sm

linreg = (sm.GLM(endog = possum['head_l'],
                 exog = sm.add_constant(possum[[]]) ,
                 family = sm.families.Gaussian())
          .fit()
         )
```

We are not going to use any other variables in our initial model. This looks strange now, but will make sense once we add a predictor.

Approach 1: Ignore the Total Length variable, and just look at the overall distribution of Head Length.

```
import statsmodels.api as sm

linreg = (sm.GLM(endog = possum['head_l'],
                 exog = sm.add_constant(possum[[]]),
                 family = sm.families.Gaussian() )
           .fit()
          )
```

We'll assume that the target follows a Gaussian (normal) distribution.

Approach 1: Ignore the Total Length variable, and just look at the overall distribution of Head Length.

```
import statsmodels.api as sm

linreg = (sm.GLM(endog = possum['head_l'],
                 exog = sm.add_constant(possum[[]]),
                 family = sm.families.Gaussian())
          .fit()
         )
```

Go ahead and fit the model after specifying it.

Approach 1: Ignore the Total Length variable, and just look at the overall distribution of Head Length.

```
linreg.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	head_l	**No. Observations:**	104
Model:	GLM	**Df Residuals:**	103
Model Family:	Gaussian	**Df Model:**	0
Link Function:	identity	**Scale:**	12.769
Method:	IRLS	**Log-Likelihood:**	-279.51
Date:	Wed, 15 Sep 2021	**Deviance:**	1315.2
Time:	17:09:34	**Pearson chi2:**	1.32e+03
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	92.6029	0.350	264.281	0.000	91.916	93.290

Approach 1: Ignore the Total Length variable, and just look at the overall distribution of Head Length.

```
linreg.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	head_l	**No. Observations:**	104
Model:	GLM	**Df Residuals:**	103
Model Family:	Gaussian	**Df Model:**	0
Link Function:	identity	**Scale:**	12.769
Method:	IRLS	**Log-Likelihood:**	-279.51
Date:	Wed, 15 Sep 2021	**Deviance:**	1315.2
Time:	17:09:34	**Pearson chi2:**	1.32e+03
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	92.6029	0.350	264.281	0.000	91.916	93.290

The estimated mean of the distribution of head lengths is 92.6029.

Approach 1: Ignore the Total Length variable, and just look at the overall distribution of Head Length.

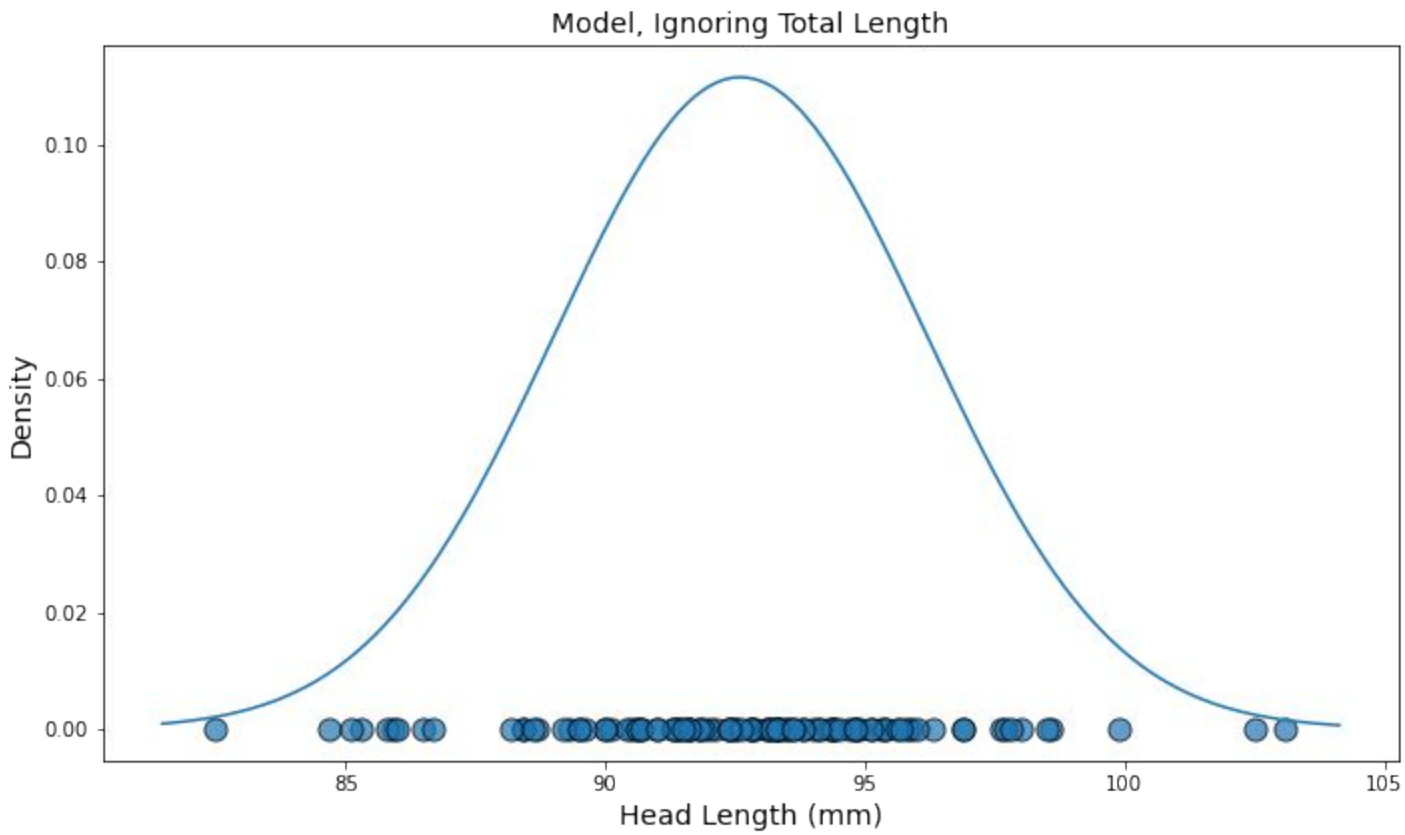
```
linreg.summary()
```

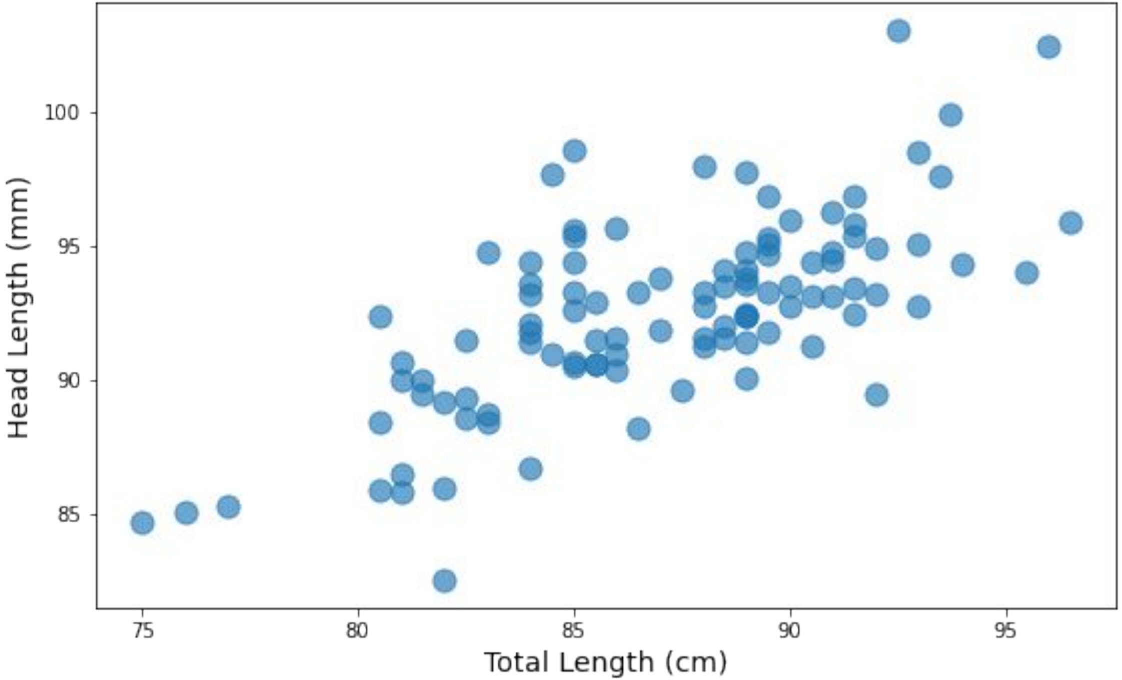
Generalized Linear Model Regression Results

Dep. Variable:	head_l	**No. Observations:**	104
Model:	GLM	**Df Residuals:**	103
Model Family:	Gaussian	**Df Model:**	0
Link Function:	identity	**Scale:**	12.769
Method:	IRLS	**Log-Likelihood:**	-279.51
Date:	Wed, 15 Sep 2021	**Deviance:**	1315.2
Time:	17:09:34	**Pearson chi2:**	1.32e+03
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	92.6029	0.350	264.281	0.000	91.916	93.290

The estimated variance of the distribution of head lengths is 12.769.


Model, Ignoring Total Length
Density
0.00
0.02
0.04
0.06
0.08
0.10
85
90
95
100
105
Head Length (mm)



The results from approach 1 look *okay*, but we are disregarding a lot of potentially useful information - the total length measurement.

Approach 2: Predict using the total length (and a constant).

Approach 2: Predict using the total length (and a constant).

```
linreg_tl = (sm.GLM(endog = possum['head_l'],
                    exog =
sm.add_constant(possum[['total_l']]),
                    family = sm.families.Gaussian())
            .fit()
           )
```

Approach 2: Predict using the total length (and a constant).

```
linreg_tl = (sm.GLM(endog = possum['head_l'],
                    exog =
sm.add_constant(possum[['total_l']]) ,
                    family = sm.families.Gaussian())
           .fit()
          )
```

This time, we'll use the total length column as a predictor.

Approach 2: Predict using the total length (and a constant).

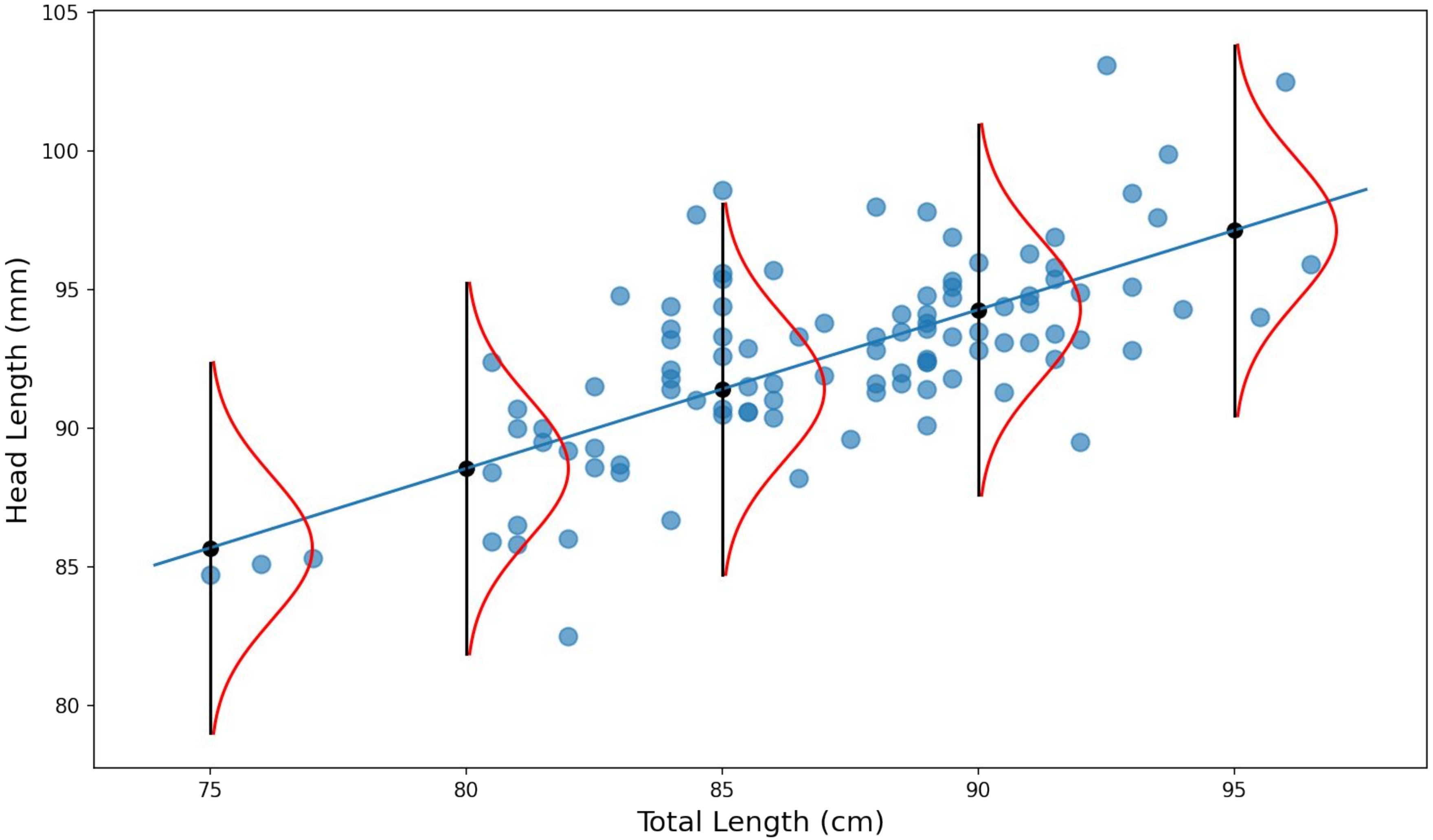
```
linreg_tl.summary()
```

Generalized Linear Model Regression Results

Dep. Variable:	head_l	**No. Observations:**	104
Model:	GLM	**Df Residuals:**	102
Model Family:	Gaussian	**Df Model:**	1
Link Function:	identity	**Scale:**	6.7357
Method:	IRLS	**Log-Likelihood:**	-245.75
Date:	Wed, 15 Sep 2021	**Deviance:**	687.04
Time:	22:16:23	**Pearson chi2:**	687.
No. Iterations:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	42.7098	5.173	8.257	0.000	32.571	52.848
total_l	0.5729	0.059	9.657	0.000	0.457	0.689

For possums with a total length of t, the model estimates that the distribution of head lengths is normal with a mean of $42.7098 + 0.5729t$ and a variance of 6.7357.



105
100
95
90
85
80
Head Length (mm)
75
80
85
90
95
Total Length (cm)

Linear Regression

We have estimated the distribution of head lengths, *conditional* on the total length.

Linear Regression

We have estimated the distribution of head lengths, *conditional* on the total length.

If we let Y be the head length and x be the total length, we have estimated the distribution of $Y|x$.

Linear Regression

We have estimated the distribution of head lengths, *conditional* on the total length.

If we let Y be the head length and x be the total length, we have estimated the distribution of $Y|x$.

Specifically, we have said that it follows a normal distribution with mean $42.7098 + 0.5729x$

Linear Regression in General

$Y|x$ follows a normal distribution with mean

$$\mu = \beta_0 + \beta_1 x$$

Linear Regression

What if we have more predictors?

Linear Regression

What if we have more predictors?

For example, along with total length x_1, we could include skull width as x_2.

Linear Regression

What if we have more predictors?

For example, along with total length x_1, we could include skull width as x_2.

If we do, we'd estimate that $Y|(x_1, x_2)$ is normal with mean

$$\mu = 29.6127 + 0.3634x_1 + 0.551x_2$$

Linear Regression in General

$Y|\vec{x}$ follows a **normal** distribution with mean

$$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

Where $\vec{x} = \langle x_1, \ldots, x_n \rangle$ are the values of the predictor variables.

To Be Continued