Multiple Linear Regression

Multiple Linear Regression

Building a linear regression model using a single predictor variable is known as **simple linear regression**.

Let’s see how we can utilize more than one predictor, which is known as **multiple linear regression**.

Example: Possums

We saw how we could estimate the total length (*total_l*) of a possum using the length of its head (*head_l*).

$$(\text{total\_l})_i = 9.89 + 0.8337 \cdot (\text{head\_l})_i + \epsilon_i$$

Example: Possums

We saw how we could estimate the total length (*total_l*) of a possum using the length of its head (*head_l*).

$$(\text{total\_l})_i = 9.89 + 0.8337 \cdot (\text{head\_l})_i + \epsilon_i$$

What if we want to incorporate other information, say the length of the tail (*tail_l*)? Using a similar process to before, we can minimize the squared residuals to obtain this equation:

$$(\text{total\_l})_i = -9.84 + 0.6950 \cdot (\text{head\_l})_i + 0.880 \cdot (\text{tail\_l})_i + \epsilon_i$$

Example: Possums

$$(\text{total\_l})_i = -9.84 + 0.6950 \cdot (\text{head\_l})_i + 0.880 \cdot (\text{tail\_l})_i + \epsilon_i$$

The change in average total length for a one-unit increase in head length, **if tail length is held constant.**

Example: Possums

$$(\text{total\_l})_i = -9.84 + 0.6950 \cdot (\text{head\_l})_i + 0.880 \cdot (\text{tail\_l})_i + \epsilon_i$$

The change in average total length for a one-unit increase in tail length, **if head length is held constant.**

Multiple Linear Regression

In general, we could use any number of predictors.

The **multiple linear regression model** is

$$Y_i = \beta_0 + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + \cdots + \beta_p \cdot X_{ip} + \epsilon_i$$

Here, we have *p* predictor variables.

The errors ε_i are assumed to be independent and normally distributed with mean of 0 and constant variance.

Multiple Linear Regression

For simple linear regression, we could ask whether the value of β_1 was statistically significant.

When doing multiple linear regression, there are many more types of questions about the coefficients we could ask.

Example: Possums Data

Let’s say we are interested in studying the relationship between a possum’s total length (*total_l*) and the following variables:

- X_1: Head length (*head_l)*
- X_2: Tail length (*tail_l*)
- X_3: Sex

Example: Possums Data

Let's say we are interested in studying the relationship between a possum's total length (*total_l*) and the following variables:

- X_1: Head length (*head_l*)
- X_2: Tail length (*tail_l*)
- X_3: Sex

Question 1: Is a regression model containing at least one predictor useful in predicting total length?

Null: H_0: $\beta_1 = \beta_2 = \beta_3 = 0$

Alternative: H_A: At least one $\beta_i \neq 0$

Example: Possums Data

Let's say we are interested in studying the relationship between a possum's total length (*total_l*) and the following variables:

- X_1: Head length (*head_l*)
- X_2: Tail length (*tail_l*)
- X_3: Sex

Question 2: Is the total length linearly related to sex, after controlling for head and tail length?

Null: $H_0: \beta_3 = 0$

Alternative: $H_A: \beta_3 \neq 0$

The General Linear F-Test

To test hypotheses about coefficients, we use the **general linear F-test**, which involves:

The General Linear F-Test

To test hypotheses about coefficients, we use the **general linear F-test**, which involves:

1. Defining a larger **full model** (which contains all predictors involved in the test)

The General Linear F-Test

To test hypotheses about coefficients, we use the **general linear F-test**, which involves:

1. Defining a larger **full model** (which contains all predictors involved in the test)
2. Defining a smaller **reduced model** (which satisfies the assumptions of the null hypothesis)

The General Linear F-Test

To test hypotheses about coefficients, we use the **general linear F-test**, which involves:

1. Defining a larger **full model** (which contains all predictors involved in the test)
2. Defining a smaller **reduced model** (which satisfies the assumptions of the null hypothesis)
3. Compare the Sum of Squares of the full and reduced model (using an F-statistic) to reach a conclusion.

The General Linear F-Test

To test hypotheses about coefficients, we use the **general linear F-test**, which involves:

1. Defining a larger **full model** (which contains all predictors involved in the test)
2. Defining a smaller **reduced model** (which satisfies the assumptions of the null hypothesis)
3. Compare the Sum of Squares of the full and reduced model (using an F-statistic) to reach a conclusion.

This can be done in *statsmodels* using the anova_lm function.

Example: Possums Data

Question 1: Is a regression model containing at least one predictor useful in predicting total length?

Null: $H_0: \beta_1 = \beta_2 = \beta_3 = 0$

Alternative: H_A: At least one $\beta_i \neq 0$

Example: Possums Data

Question 1: Is a regression model containing at least one predictor useful in predicting total length?

Null: $H_0: \beta_1 = \beta_2 = \beta_3 = 0$

Alternative: H_A: At least one $\beta_i \neq 0$

```
lr_reduced = sm.ols('total_l ~ 1', data = possum).fit()
lr_full = sm.ols('total_l ~ head_l + tail_l + sex', data = possum).fit()
```

Example: Possums Data

Question 1: Is a regression model containing at least one predictor useful in predicting total length?

Null: $H_0: \beta_1 = \beta_2 = \beta_3 = 0$

Alternative: H_A: At least one $\beta_i \neq 0$

```
lr_reduced = sm.ols('total_l ~ 1', data = possum).fit()
lr_full = sm.ols('total_l ~ head_l + tail_l + sex', data = possum).fit()

stats.stats.anova_lm(lr_reduced, lr_full)
```

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	103.0	1913.826154	0.0	NaN	NaN	NaN
1	100.0	636.471605	3.0	1277.354549	66.897698	8.190568e-24

Conclusion: reject the null hypothesis

Example: Possums Data

Question 2: Is the total length linearly related to sex, after controlling for head and tail length?

Null: $H_0: \beta_3 = 0$

Alternative: $H_A: \beta_3 \neq 0$

Example: Possums Data

Question 2: Is the total length linearly related to sex, after controlling for head and tail length?

Null: $H_0: \beta_3 = 0$

Alternative: $H_A: \beta_3 \neq 0$

```
lr_reduced = sm.ols('total_l ~ head_l + tail_l', data = possum).fit()
lr_full = sm.ols('total_l ~ head_l + tail_l + sex', data = possum).fit()
```

Example: Possums Data

Question 2: Is the total length linearly related to sex, after controlling for head and tail length?

Null: $H_0: \beta_3 = 0$

Alternative: $H_A: \beta_3 \neq 0$

```
lr_reduced = sm.ols('total_l ~ head_l + tail_l', data = possum).fit()
lr_full = sm.ols('total_l ~ head_l + tail_l + sex', data = possum).fit()

stats.stats.anova_lm(lr_reduced, lr_full)
```

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	101.0	718.770150	0.0	NaN	NaN	NaN
1	100.0	636.471605	1.0	82.298545	12.930435	0.000504

Conclusion: reject the null hypothesis

Example: Possums Data

What if we fit all of the variables?

```
lr_full = sm.ols('total_l ~ pop + sex + age + head_l + skull_w + tail_l', data = possum).fit()
```

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-14.8946	6.620	-2.250	0.027	-28.036	-1.753
pop[T.other]	-3.1541	0.536	-5.886	0.000	-4.218	-2.090
sex[T.m]	-0.9432	0.467	-2.021	0.046	-1.870	-0.016
age	-0.0421	0.122	-0.344	0.731	-0.285	0.201
head_l	0.5746	0.091	6.305	0.000	0.394	0.756
skull_w	0.0727	0.100	0.729	0.468	-0.125	0.271
tail_l	1.2749	0.138	9.260	0.000	1.002	1.548

Example: Possums Data

What if we fit all of the variables?

```
lr_full = sm.ols('total_l ~ pop + sex + age + head_l + skull_w + tail_l', data = possum).fit()
```

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-14.8946	6.620	-2.250	0.027	-28.036	-1.753
pop[T.other]	-3.1541	0.536	-5.886	0.000	-4.218	-2.090
sex[T.m]	-0.9432	0.467	-2.021	0.046	-1.870	-0.016
age	-0.0421	0.122	-0.344	0.731	-0.285	0.201
head_l	0.5746	0.091	6.305	0.000	0.394	0.756
skull_w	0.0727	0.100	0.729	0.468	-0.125	0.271
tail_l	1.2749	0.138	9.260	0.000	1.002	1.548

We end up with two variables that are not statistically significant on an individual level. But these individual tests are affected by correlations among variables, so let's use the general linear F test to determine if we can drop both.

Null: $H_0: \beta_{age} = \beta_{skull\_w} = 0$

Alternative: $H_A: \beta_{age} \neq 0$ or $\beta_{skull\_w} \neq 0$

Example: Possums Data

Null: $H_0: \beta_{age} = \beta_{skull\_w} = 0$

Alternative: $H_A: \beta_{age} \neq 0$ or $\beta_{skull\_w} \neq 0$

```
lr_full = sm.ols('total_l ~ pop + sex + age + head_l + skull_w + tail_l', data = possum).fit()
lr_reduced = sm.ols('total_l ~ pop + sex + head_l + tail_l', data = possum).fit()
```

Example: Possums Data

Null: $H_0: \beta_{age} = \beta_{skull\_w} = 0$

Alternative: $H_A: \beta_{age} \neq 0$ or $\beta_{skull\_w} \neq 0$

```
lr_full = sm.ols('total_l ~ pop + sex + age + head_l + skull_w + tail_l', data = possum).fit()
lr_reduced = sm.ols('total_l ~ pop + sex + head_l + tail_l', data = possum).fit()

stats.stats.anova_lm(lr_reduced, lr_full)
```

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	99.0	492.111692	0.0	NaN	NaN	NaN
1	95.0	455.509768	4.0	36.601924	1.908402	0.115344

Conclusion: do not reject the null hypothesis.
We can drop age and skull_w from the model.

Interaction Terms

Let’s see what happens if we fit a model using head_l and sex.

	coef
Intercept	8.2610
sex[T.m]	-2.0646
head_l	0.8643

Interaction Terms

Let's see what happens if we fit a model using head_l and sex.

	coef
Intercept	8.2610
sex[T.m]	-2.0646
head_l	0.8643

For male possums:

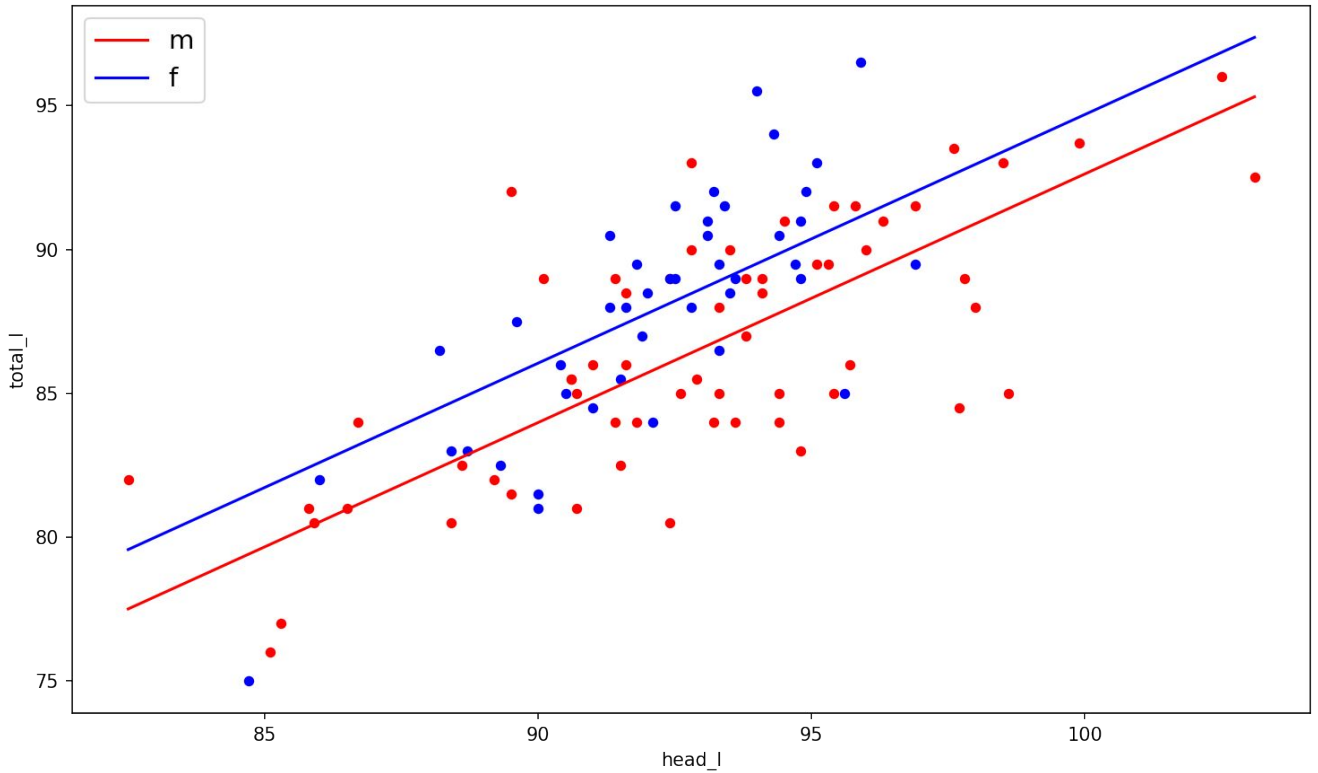
$$(\text{total\_l})_i = 8.261 - 2.0646 + 0.8643 \cdot (\text{head\_l})_i + \epsilon_i$$
$$(\text{total\_l})_i = 6.1964 + 0.8643 \cdot (\text{head\_l})_i + \epsilon_i$$

For female possums:

$$(\text{total\_l})_i = 8.261 + 0.8643 \cdot (\text{head\_l})_i + \epsilon_i$$

This gives parallel lines, implying that the effect of head_l is the same, regardless of sex.

Interaction Terms



Interaction Terms

If we want to test whether the effect of head_l is different for male and female possums, we can use **interaction** terms, which we get by multiplying two variables together.

Interaction Terms

If we want to test whether the effect of head_l is different for male and female possums, we can use **interaction** terms, which we get by multiplying two variables together.

	coef
Intercept	-28.7222
sex[T.m]	45.0841
head_l	1.2657
head_l:sex[T.m]	-0.5107

Interaction Terms

If we want to test whether the effect of head_l is different for male and female possums, we can use **interaction** terms, which we get by multiplying two variables together.

	coef
Intercept	-28.7222
sex[T.m]	45.0841
head_l	1.2657
head_l:sex[T.m]	-0.5107

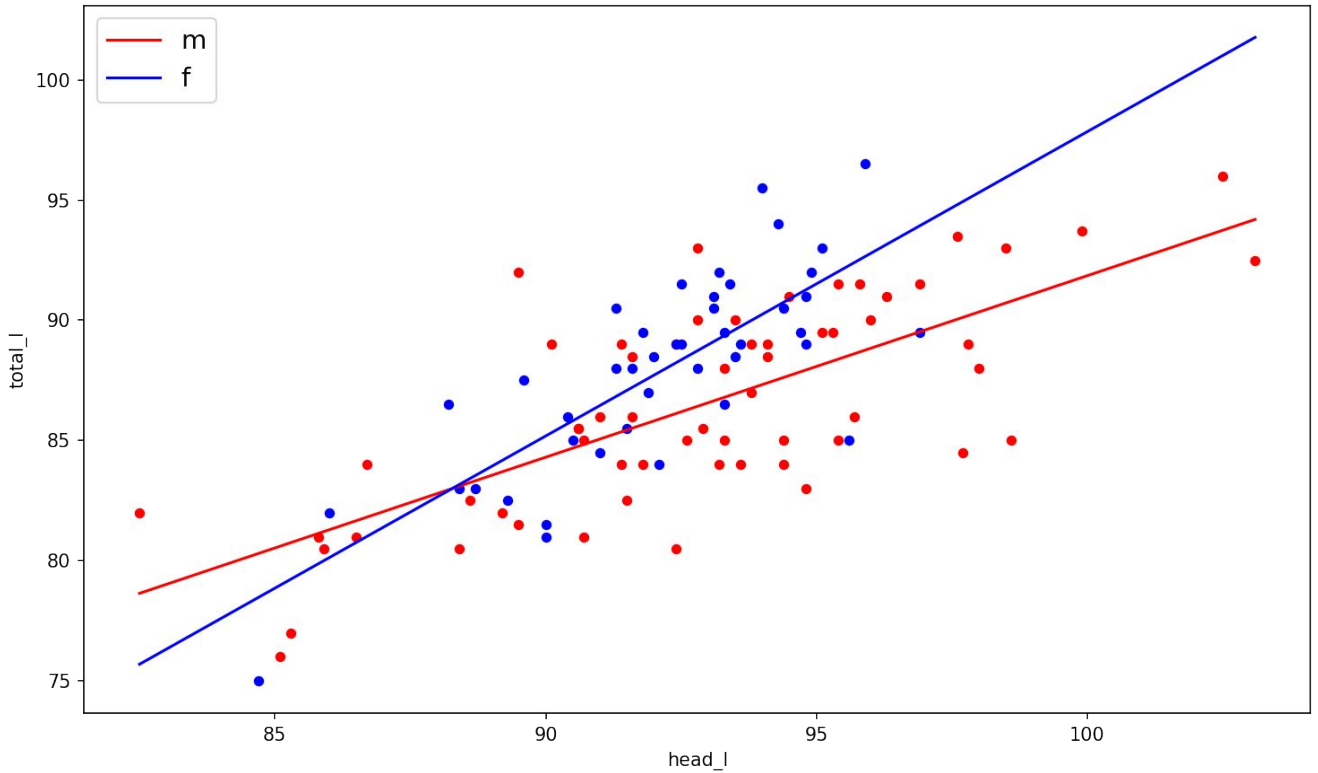
For male possums:

$$(\text{total\_l})_i = 16.3621 + 0.755 \cdot (\text{head\_l})_i + \epsilon_i$$

For female possums:

$$(\text{total\_l})_i = -28.7222 + 1.2657 \cdot (\text{head\_l})_i + \epsilon_i$$

Interaction Terms



Interaction Terms

Question: Are these interaction terms statistically significant?

	coef
Intercept	-28.7222
sex[T.m]	45.0841
head_l	1.2657
head_l:sex[T.m]	-0.5107

Interaction Terms

Question: Are these interaction terms statistically significant?

Null: $H_0: \beta_{head\_l:sex} = 0$

Alternative: $H_A: \beta_{head\_l:sex} \neq 0$

	coef
Intercept	-28.7222
sex[T.m]	45.0841
head_l	1.2657
head_l:sex[T.m]	-0.5107

Interaction Terms

Question: Are these interaction terms statistically significant?

Null: $H_0: \beta_{head\_l:sex} = 0$

Alternative: $H_A: \beta_{head\_l:sex} \neq 0$

	coef
Intercept	-28.7222
sex[T.m]	45.0841
head_l	1.2657
head_l:sex[T.m]	-0.5107

```
lr_full =sm.ols('total_l ~ head_l + sex + head_l:sex', data = possum).fit()
lr_reduced =sm.ols('total_l ~ head_l + sex', data = possum).fit()
```

Interaction Terms

Question: Are these interaction terms statistically significant?

Null: $H_0: \beta_{head\_l:sex} = 0$

Alternative: $H_A: \beta_{head\_l:sex} \neq 0$

	coef
Intercept	-28.7222
sex[T.m]	45.0841
head_l	1.2657
head_l:sex[T.m]	-0.5107

```
lr_full =sm.ols('total_l ~ head_l + sex + head_l:sex', data = possum).fit()
lr_reduced =sm.ols('total_l ~ head_l + sex', data = possum).fit()

stats.stats.anova_lm(lr_reduced, lr_full)
```

	df_resid	ssr	df_diff	ss_diff	F	Pr(>F)
0	101.0	893.49296	0.0	NaN	NaN	NaN
1	100.0	836.41236	1.0	57.0806	6.824457	0.010377

Conclusion: reject the null hypothesis. The interaction term is significant.

Multiple Linear Regression - Other Considerations

When doing inference with multiple regression, it is still important to check the assumptions.

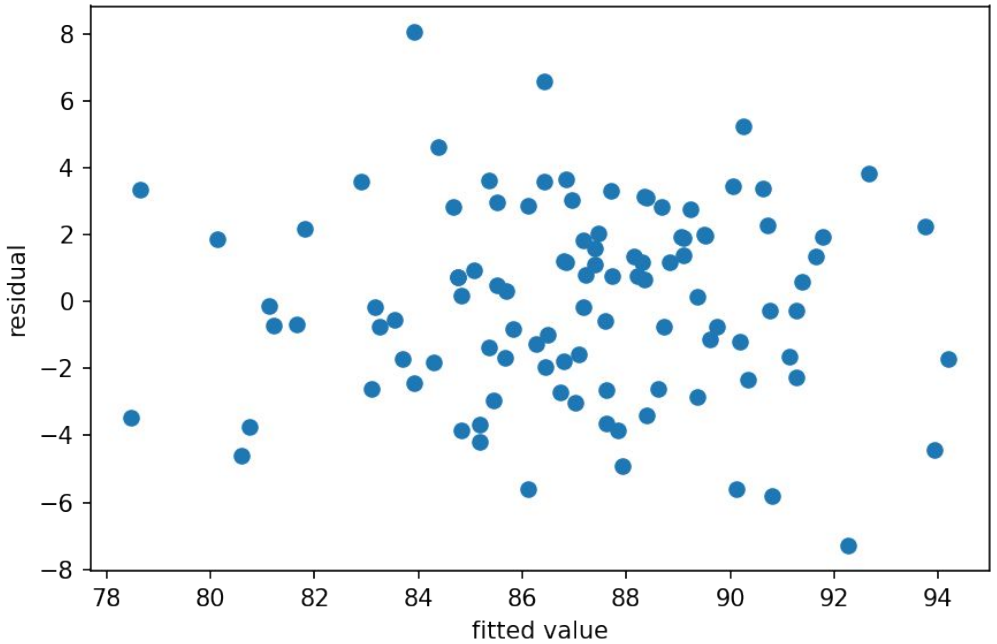
This is usually done by inspecting plots of the residuals.

We want to see normality and constant variance with no patterns.

Types of scatterplots to consider:

- residuals vs. fitted values
- residuals vs. predictor
- residuals vs. other variables not included

Multiple Linear Regression - Other Considerations

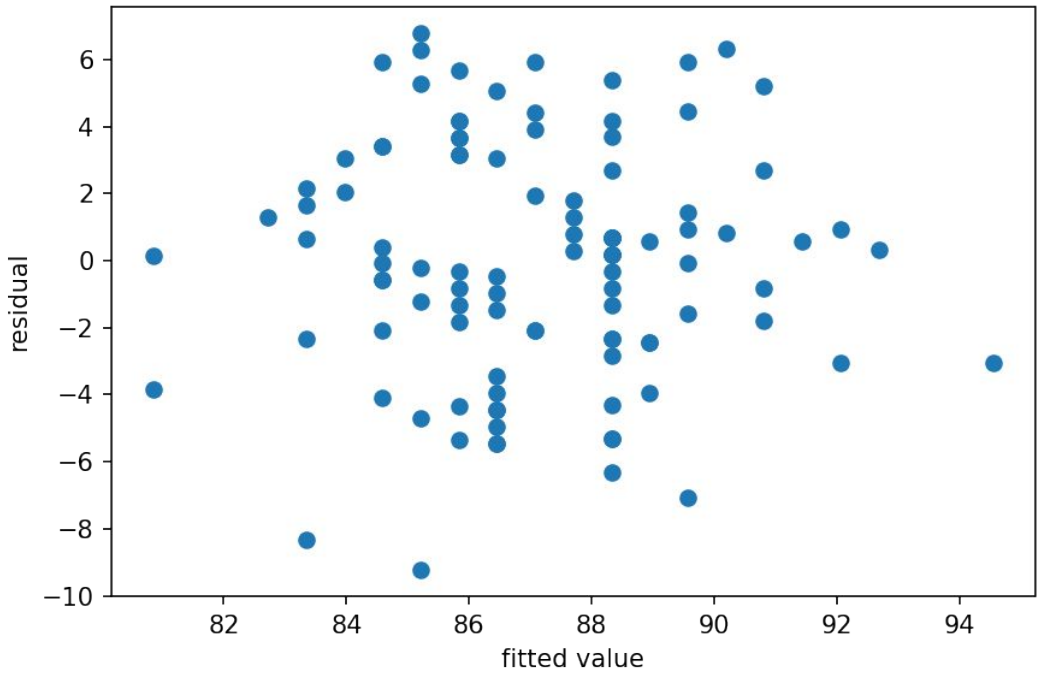


There is no obvious pattern in these residuals.

Multiple Linear Regression - Other Considerations

Let’s fit a model for total_l using tail_l alone.

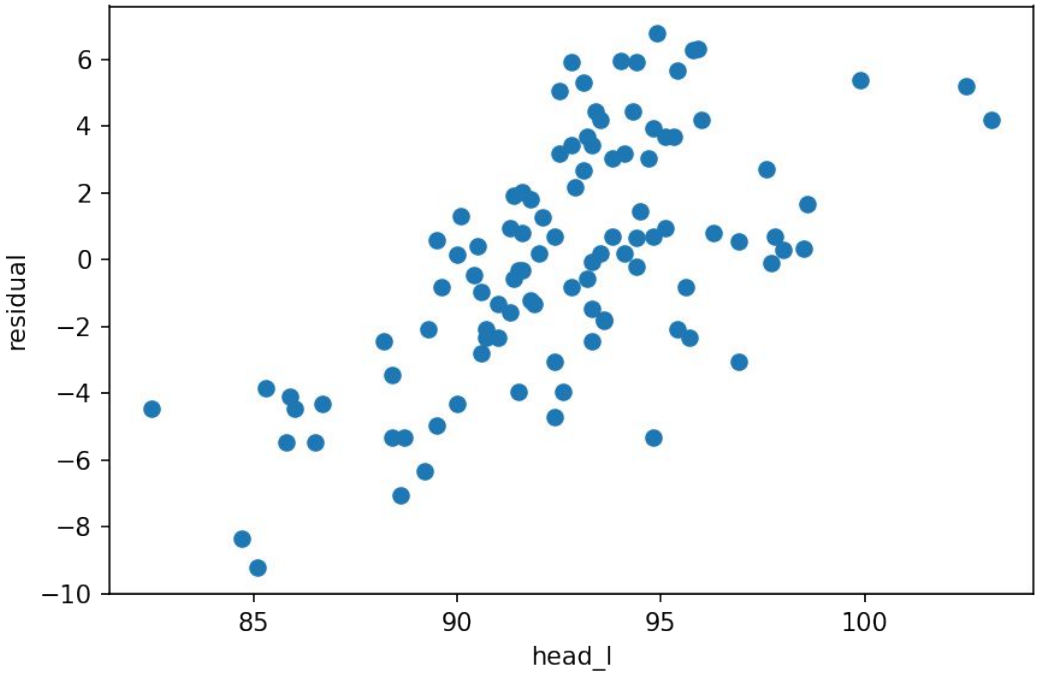
We can look at the residuals vs. fitted values. There are not obvious patterns, but the distribution looks potentially skewed.



Multiple Linear Regression - Other Considerations

Let's fit a model for total_l using tail_l alone.

If we look at residuals vs. head_l, we can see an obvious pattern. This can indicate that we would benefit from adding this variable to our model.



Multiple Linear Regression - Multicollinearity

If we look at the coefficient for head_l, it changes by quite a bit when we also include tail_l in our model.

$$(\text{total\_l})_i = 9.89 + 0.8337 \cdot (\text{head\_l})_i + \epsilon_i$$

$$(\text{total\_l})_i = -9.84 + 0.6950 \cdot (\text{head\_l})_i + 0.880 \cdot (\text{tail\_l})_i + \epsilon_i$$

Why? Because these variables are moderately correlated.

Multicollinearity exists when two or more predictors are moderately or highly correlated.

Multiple Linear Regression - Multicollinearity

Effects of multicollinearity:

- Estimated coefficients depend on which variables are included in the model (and can sometimes be counter to what might be expected - negative instead of positive, for example).
- Precision of estimates decreases.
- The results of testing some β_i may change, depending on which variables are included.
- Estimates of mean response **is not** affected.

Multiple Linear Regression in Python

Let’s see some of this in action in the notebook.