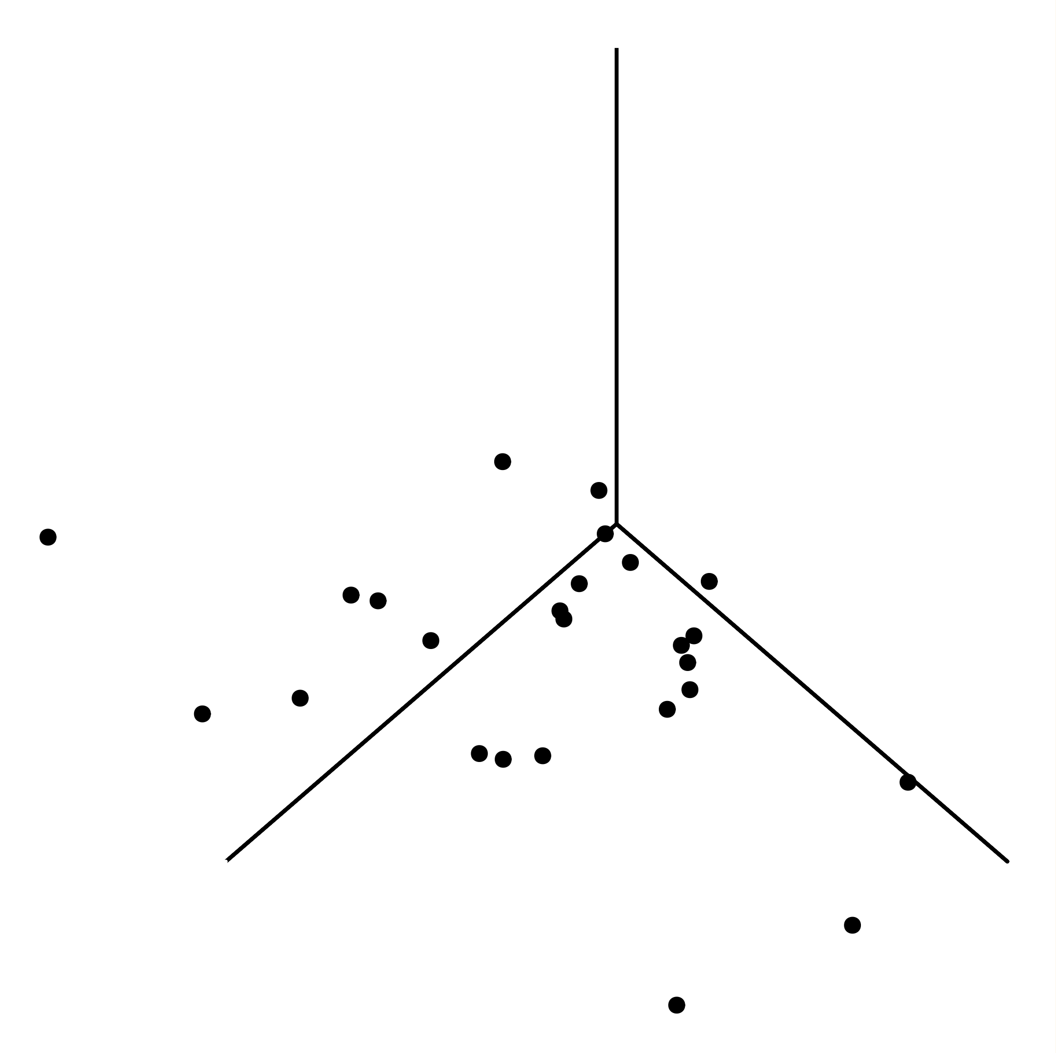
MLR Inference

Stat 230

April 18 2022

Overview



Today:

MLR Inference

- t tests/CIs for coefficients
- CI for mean response, PI for new response
- Linear combinations of coefficients: CIs and tests

MLR Model

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p + \epsilon_i, \epsilon_i \sim N(0, \sigma)$$

MLR inference:

- estimation is the same as SLR!
- tests/CI are also same as SLR!
  - **Except** we use $n - (p + 1)$ degrees of freedom where $p + 1$ is the number of β's in the model (including the intercept)

MLR inference for one parameter: Testing

- Testing single parameters:

$$H_0 : \beta_j = 0 \quad H_A : \beta_j \neq 0$$
$$t = \frac{\hat{\beta}_j - 0}{SE\left(\hat{\beta}_j\right)}$$
$$\text{p-value } = 2 \times P(T > |t|)$$

using t-distribution with $n - (p+1)$ degrees of freedom

MLR inference for one parameter: Confidence intervals (CI)

- CI for a single parameters:

$$\hat{\beta}_j \pm t^* SE\left(\hat{\beta}_j\right)$$

t^* is the $(100 - C)/2$ percentile from the t-distribution with $n - (p + 1)$ degrees of freedom

Example: RECS MLR

```r
library(moderndive)
energy <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/EnergySurvey
energy <- energy %>%
  mutate(logCost = log(CostTotal),
         logSqft = log(SqftMeasure))

cost_lm_log <- lm(logCost ~ logSqft + HHSize, data = energy)
reg_table_energy <- get_regression_table(cost_lm_log, digits = 4)
knitr::kable(reg_table_energy, format = "html")
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	4.3667	0.0711	61.3860	0	4.2272	4.5061
logSqft	0.3722	0.0096	38.7495	0	0.3533	0.3910
HHSize	0.0839	0.0046	18.1906	0	0.0748	0.0929

$$\hat{\mu}(\log(\text{ Cost }) \mid x) = 4.3667 + 0.3722 \log(\text{ Sqft }) + 0.0839 \text{ HHSize}$$

Example: RECS MLR

```
get_regression_summaries(cost_lm_log) %>% knitr::kable(digits = 4)
```

r_squared	adj_r_squared	mse	rmse	sigma	statistic	p_value	df	nobs
0.329	0.329	0.1998	0.447	0.447	1075.249	0	2	4381

What does $H_0 : \beta_2 = 0$ vs. $H_A : \beta_2 \neq 0$ test?

Example: RECS MLR

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
HHSize	0.0839	0.0046	18.1906	0	0.0748	0.0929

r_squared	adj_r_squared	mse	rmse	sigma	statistic	p_value	df	nobs
0.329	0.329	0.1998	0.447	0.447	1075.249	0	2	4381

- Is household size associated with energy cost after controlling for dwelling size?
- Test stat? p-value?

$$t = \frac{0.0839 - 0}{0.0046} \approx 18.19$$

$$\text{p-value} \; = 2 \times P(T > 18.19) = 2 \times P(T < -18.19) = 2 \times \text{pt}(-18.19, \text{df} = 4381 - 3) < 0.0001$$

```
# 2 times area in the left tail
# beyond negative observed test stat
2*pt(-18.19, df = 4378)
[1] 2.473683e-71
```

```
# 2 times area in the right tail
# beyond the observed test stat
2*(1 - pt(18.19, df = 4378))
[1] 0
```

Example: RECS MLR

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
HHSize	0.0839	0.0046	18.1906	0	0.0748	0.0929

Is household size associated with energy cost after controlling for dwelling size?

- Yes, after accounting for dwelling square footage, the estimated effect of household size on mean energy costs is statistically significant $(\mathrm{t} = 18.19, \mathrm{df} = 4378, \mathrm{p} < 0.0001)$.
- Quantify the "effect": 95% CI for β_2

$$0.0839 \pm (1.9605)(0.0046) = (0.075, 0.093)$$

```
qt(.975, df = 4378)
[1] 1.960506
0.0839 + c(-1,1)*(1.9605)*(0.0046)
[1] 0.0748817 0.0929183
```

Example: RECS MLR

- Quantify the "effect": 95% CI for β_2 is 0.075 to 0.093
- Our response was logged but HHSize was not (think, SLR exponential model)

```
cost_lm_log <- lm(logCost ~ logSqft + HHSize, data = energy)
```

A 1 person increase in household size is associated with a multiplicative change in median energy cost between

```
exp(c(.075, .093))
[1] 1.077884 1.097462
```

We are 95% confident that a 1 person increase in household size is associated with 7.9% to 9.7% increase in median energy cost after accounting for dwelling size.

Example: Pengiuns from `palmerpenguins`

Model mean mass as a function of bill length, species and their interaction:

$$\begin{aligned}\mu_{\text{mass}\,|x} &= \beta_0 + \beta_1 \text{ bill } + \beta_2 \text{ speciesChinstrap } + \beta_3 \text{ speciesGentoo} \\ &+ \beta_4 \text{ bill } \times \text{ speciesChinstrap } + \beta_5 \text{ bill } \times \text{ speciesGentoo}\end{aligned}$$

- Mean function for Adelie (baseline species)

$$\begin{aligned}\mu_{\text{mass}\,|x} &= \beta_0 + \beta_1 \text{ bill } + \beta_2(0) + \beta_3(0) + \beta_4 bill \times (0) + \beta_5 bill \times (0) \\ &= \beta_0 + \beta_1 \text{ bill}\end{aligned}$$

- β_1 : effect of bill length on mass for Adelie (baseline)
- 95% CI for this parameter??

Example: Pengiuns from palmerpenguins

```r
# load library and remove NA cases for the 3 variables of interest
library(palmerpenguins)
library(tidyr) # has `drop_na()` function that drops rows having missing values
penguins <- penguins %>% tidyr::drop_na(bill_length_mm, body_mass_g, species)
peng_interaction_lm <-  lm(body_mass_g ~ bill_length_mm*species, data = penguins)
peng_table_interaction <- get_regression_table(peng_interaction_lm, digits = 5)
knitr::kable(peng_table_interaction, digis= 5, format = "html")
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	34.88299	443.17604	0.07871	0.93731	-836.86616	906.63214
bill_length_mm	94.49982	11.39794	8.29095	0.00000	72.07950	116.92013
species: Chinstrap	811.26034	799.80552	1.01432	0.31116	-761.99661	2384.51729
species: Gentoo	-158.71092	683.19141	-0.23231	0.81644	-1502.58216	1185.16031
bill_length_mm:speciesChinstrap	-35.38208	17.74666	-1.99373	0.04699	-70.29064	-0.47353
bill_length_mm:speciesGentoo	14.95935	15.78642	0.94761	0.34401	-16.09333	46.01202

Example: Pengiuns from `palmerpenguins`

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
bill_length_mm	94.49982	11.39794	8.29095	0	72.0795	116.9201

r_squared	adj_r_squared	mse	rmse	sigma	statistic	p_value	df	nobs
0.788	0.785	135810	368.524	371.8	250.097	0	5	342

```
qt(.975, df = 336) # df = 342 - 6 = 336
[1] 1.967049
```

- A one mm increase in bill length in Adelie penguins is associated with a $72.1\ \mathrm{g}$ to $116.9\ \mathrm{g}$ increase in mean weight.

$$94.49982 \pm (1.9670)(11.39794) = 72.08, 116.92$$

Example: Pengiuns from `palmerpenguins`

Model mean mass as a function of bill length, species and their interaction:

$$\begin{aligned}\mu_{\text{mass }|x} &= \beta_0 + \beta_1 \text{ bill } + \beta_2 \text{ speciesChinstrap } + \beta_3 \text{ speciesGentoo} \\ &+ \beta_4 \text{ bill } \times \text{ speciesChinstrap } + \beta_5 \text{ bill } \times \text{ speciesGentoo}\end{aligned}$$

- Mean function for Chinstrap

$$\begin{aligned}\mu_{\text{mass }|x} &= \beta_0 + \beta_1 bill + \beta_2(1) + \beta_3(0) + \beta_4 bill \times (1) + \beta_5 bill \times (0) \\ &= \beta_0 + \beta_2 + (\beta_1 + \beta_4)\, bill\end{aligned}$$

- $\beta_1 + \beta_4$: effect of bill length on mass for Chinstrap
- 95% CI for this parameter??

Sampling distribution of $\hat{\beta}_j$'s

Before, we looked at how $\hat{\beta}_j$ estimates are correlated

- For penguins:
- $\hat{\beta}_1$ and $\hat{\beta}_4$ are also correlated, meaning

$$\boldsymbol{SE}\left(\hat{\beta}_1 + \hat{\beta}_4\right) \neq \sqrt{\boldsymbol{SE}\left(\hat{\beta}_1\right)^2 + \boldsymbol{SE}\left(\hat{\beta}_4\right)^2}$$

New Inference: linear combinations of two β_j

Parameter of interest:

$$\gamma = c_i\beta_i + c_j\beta_j$$

where c_i and c_j are known constants.

- The effect of bill length on mass for Chinstraps:

$$\gamma = (1) \times \beta_1 + (1) \times \beta_4 = \beta_1 + \beta_4$$

  where $c_1 = 1$ and $c_4 = 1$.
- Holding bill length fixed, $\beta_2 + \beta_4 bill$ is the difference in mean mass between Chinstrap and Adelie.
- If bill length is 45 mm, this difference is

$$\gamma = (1) \times \beta_2 + (45) \times \beta_4 = \beta_2 + 45\beta_4$$

  where $c_2 = 1$ and $c_4 = 45$.

New Inference: linear combinations of two β_j

Estimated parameter:

$$\gamma = c_i\beta_i + c_j\beta_j$$

where c_i and c_j are known constants.

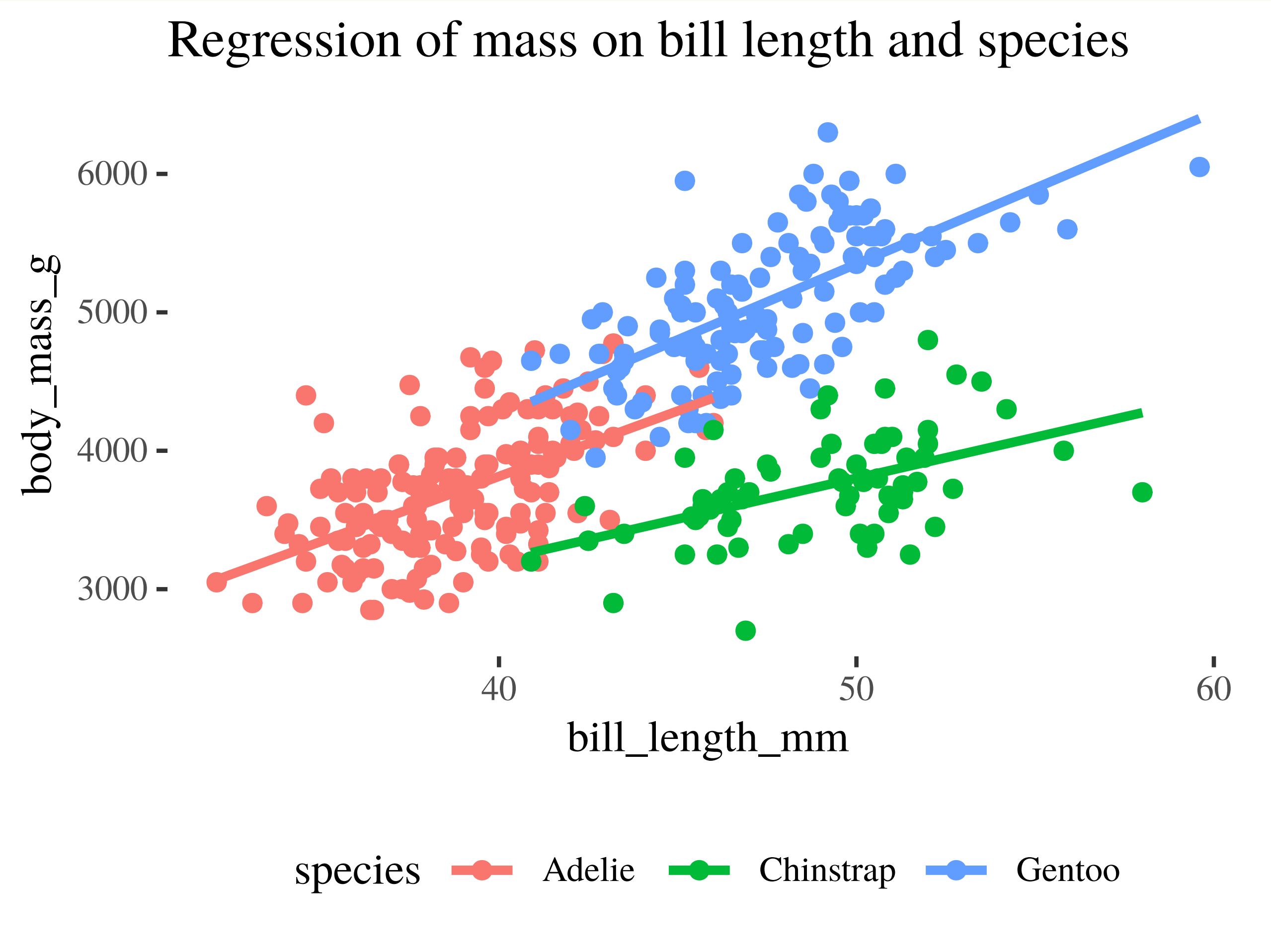
```r
knitr::kable(peng_table_interaction[c(2,5),], digits= 5, format = "html")
```

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
bill_length_mm	94.49982	11.39794	8.29095	0.00000	72.07950	116.92013
bill_length_mm:speciesChinstrap	-35.38208	17.74666	-1.99373	0.04699	-70.29064	-0.47353

- The estimated effect of bill length on mass for Chinstraps:

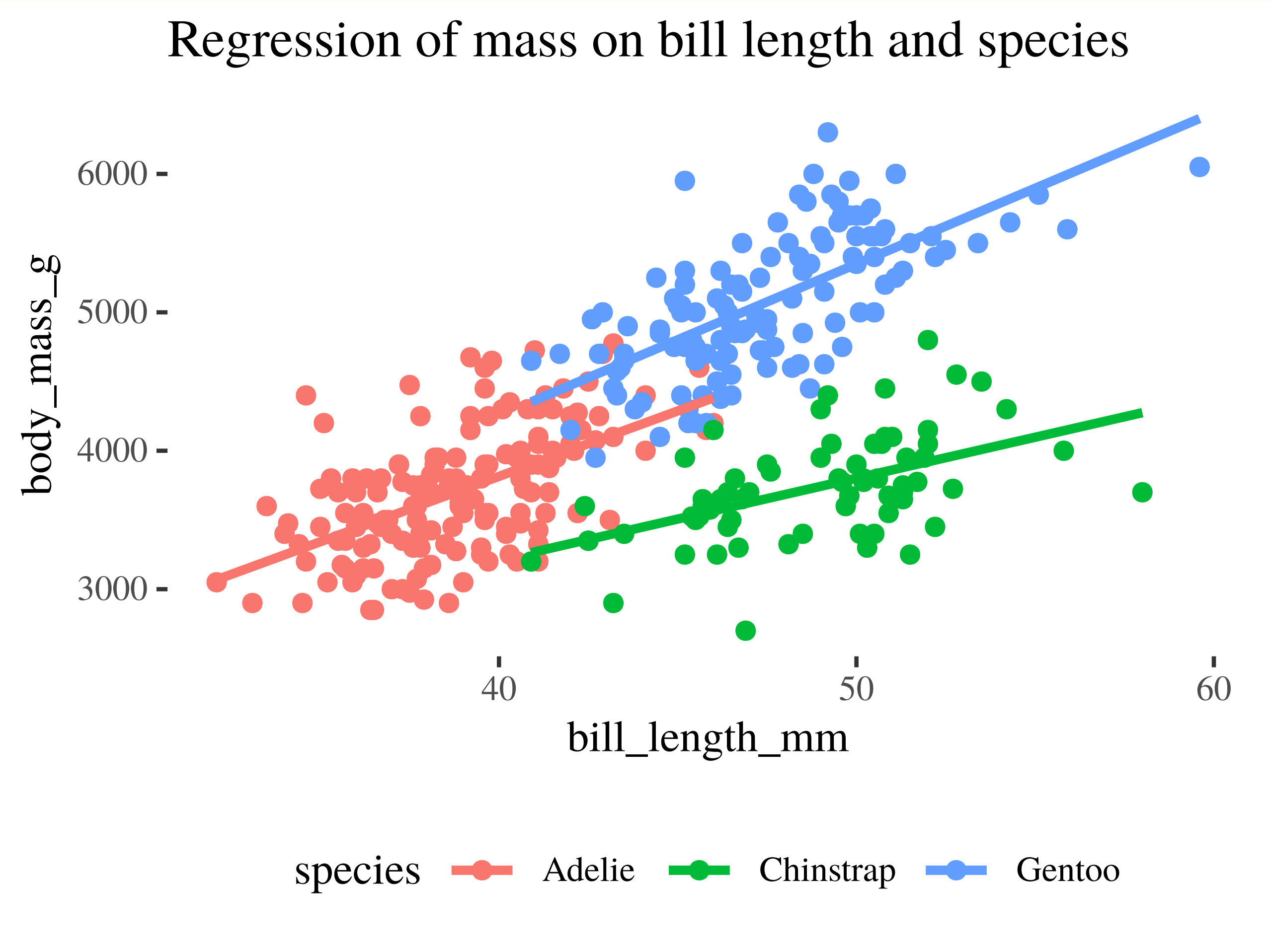
$$\hat{\gamma} = \hat{\beta}_1 + \hat{\beta}_4 = 94.49982 + (-35.38208) = 59.11774$$

New Inference: linear combinations of two β_j



Quick check: does it make sense that the effect of bill length on mass (**slope**) less for Chinstrap than Adelie?

New Inference: linear combinations of two β_j



Quick check: does it make sense that the effect of bill length on mass (**slope**) less for Chinstrap than Adelie?

- **Yes:** The green line (chinstrap) is a bit flatter than the red line (adelie)

New Inference: linear combinations of two β_j

SE of the Estimated parameter:

$$SE(\hat{\gamma}) = \sqrt{c_i^2 \operatorname{Var}\left(\hat{\beta}_i\right) + c_j^2 \operatorname{Var}\left(\hat{\beta}_j\right) + 2c_i c_j \operatorname{Cov}\left(\hat{\beta}_i, \hat{\beta}_j\right)}$$

where c_i and c_j are known constants.

- "Var" is the variance of an estimator, which equals

$$\operatorname{Var}(\hat{\beta}) = SE(\hat{\beta})^2$$

- "Cov" is the covariance between two estimators: how do they "co-vary" together?
- an "unscaled" version of correlation

$$\operatorname{Cov}\left(\hat{\beta}_i, \hat{\beta}_j\right) = \text{ how do } \hat{\beta}_i \text{ and } \hat{\beta}_j \text{ co-vary?}$$

New Inference: linear combinations of two β_j

SE of the Estimated parameter:

$$SE(\hat{\gamma}) = \sqrt{c_i^2 \operatorname{Var}\left(\hat{\beta}_i\right) + c_j^2 \operatorname{Var}\left(\hat{\beta}_j\right) + 2c_ic_j \operatorname{Cov}\left(\hat{\beta}_i, \hat{\beta}_j\right)}$$

where c_i and c_j are known constants.

- The effect of bill length on mass for Chinstraps:

$$\gamma = (1) \times \beta_1 + (1) \times \beta_4 = \beta_1 + \beta_4$$

  where $c_1 = 1$ and $c_4 = 1$.

- The SE of the estimated effect of bill length on mass for Chinstraps:

$$SE(\hat{\gamma}) = \sqrt{1^2 \operatorname{Var}\left(\hat{\beta}_1\right) + 1^2 \operatorname{Var}\left(\hat{\beta}_4\right) + 2(1)(1) \operatorname{Cov}\left(\hat{\beta}_1, \hat{\beta}_4\right)}$$

New Inference: linear combinations of two β_j

- The SE of the estimated effect of bill length on mass for Chinstraps:

$$SE(\hat{\gamma}) = \sqrt{1^2 \operatorname{Var}\left(\hat{\beta}_1\right) + 1^2 \operatorname{Var}\left(\hat{\beta}_4\right) + 2(1)(1) \operatorname{Cov}\left(\hat{\beta}_1, \hat{\beta}_4\right)}$$

- Both $\operatorname{Var}\left(\hat{\beta}_j\right)$'s and $\operatorname{Cov}\left(\hat{\beta}_i, \hat{\beta}_j\right)$ are obtained using the `vcov` command
- `vcov` returns a $(p+1) \times (p+1)$ matrix (rows/cols for each $\hat{\beta}_j$ in a model)
- diagonal values are $\operatorname{Var}\left(\hat{\beta}_j\right)$'s
- off-diagonal values are $\operatorname{Cov}\left(\hat{\beta}_i, \hat{\beta}_j\right)$'s
- row/column combination determines "i" and "j"

New Inference: linear combinations of two β_j

The SE of the estimated effect of bill length on mass for Chinstraps:

$$
\begin{aligned}
SE(\hat{\gamma}) &= \sqrt{1^2 \operatorname{Var}\left(\hat{\beta}_1\right) + 1^2 \operatorname{Var}\left(\hat{\beta}_4\right) + 2(1)(1) \operatorname{Cov}\left(\hat{\beta}_1, \hat{\beta}_4\right)} \\
&= \sqrt{\operatorname{Var}\left(\hat{\beta}_1\right) + \operatorname{Var}\left(\hat{\beta}_4\right) + 2 \operatorname{Cov}\left(\hat{\beta}_1, \hat{\beta}_4\right)}
\end{aligned}
$$

The variance-covariance matrix is:

	\`$\hat{\beta}_0$\`	\`$\hat{\beta}_1$\`	\`$\hat{\beta}_2$\`	\`$\hat{\beta}_3$\`	\`$\hat{\beta}_4$\`	\`$\hat{\beta}_5$\`
\`$\hat{\beta}_0$\`	196404.999	-5039.508	-196404.999	-196404.999	5039.508	5039.508
\`$\hat{\beta}_1$\`	-5039.508	129.913	5039.508	5039.508	-129.913	-129.913
\`$\hat{\beta}_2$\`	-196404.999	5039.508	639688.864	196404.999	-14075.274	-5039.508
\`$\hat{\beta}_3$\`	-196404.999	5039.508	196404.999	466750.497	-5039.508	-10706.750
\`$\hat{\beta}_4$\`	5039.508	-129.913	-14075.274	-5039.508	314.944	129.913
\`$\hat{\beta}_5$\`	5039.508	-129.913	-5039.508	-10706.750	129.913	249.211

New Inference: linear combinations of two β_j

```r
# beta_1 = bill = row/col 2
# beta_4 = bill:chinstrap = row/col 5
vcov(peng_interaction_lm)[c(2,5), c(2,5)]
```

```
                                bill_length_mm bill_length_mm:speciesChinstrap
bill_length_mm                        129.9131                        -129.9131
bill_length_mm:speciesChinstrap      -129.9131                         314.9440
```

- $\text{Var}\left(\hat{\beta}_1\right) = 11.39794^2 = 129.9131$
- $\text{Var}\left(\hat{\beta}_4\right) = 17.74666^2 = 314.9440$
- $\text{Cov}\left(\hat{\beta}_1, \hat{\beta}_4\right) = -129.9131$

New Inference: linear combinations of two β_j

The SE of the estimated effect of bill length on mass for Chinstraps:

$$SE(\hat{\gamma}) = \sqrt{1^2 129.9131 + 1^2 314.9440 + 2(1)(1)(-129.9131)} = 13.60261$$

```
sqrt(129.9131 + 314.9440 + 2*(-129.9131))
```

```
[1] 13.60261
```

- The estimated value of $\gamma = \beta_1 + \beta_4$ is 59.11774 with a SE of 13.60261.
- Is this estimated effect of bill length on mass for Chinstraps statistically significant?
- What is a 95% CI for the true effect?

New Inference: linear combinations of two β_j

- Tests: test stat $t = \frac{\hat{\gamma} - \text{null } \gamma}{SE(\hat{\gamma})}$
- CI: $\hat{\gamma} \pm t * SE(\hat{\gamma})$
- use usual t-distribution with $\mathbf{df} = n - (p+1)$
- Is the estimated effect of bill length on mass for Chinstraps statistically significant?

$$H_0 : \gamma = 0 \quad t = \frac{59.11774 - 0}{13.60261} = 4.346$$

$$\text{p-value} = 2 \times P(T > 4.346) = 0.00002$$

```
(test_stat <- 59.11774/13.60261)
[1] 4.346059
2*pt(-4.346, df = 336)
[1] 1.83941e-05
```

New Inference: linear combinations of two β_j

- Tests: test stat $t = \frac{\hat{\gamma} - \text{null } \gamma}{SE(\hat{\gamma})}$
- CI: $\hat{\gamma} \pm t * SE(\hat{\gamma})$
- use usual t-distribution with df $= n - (p+1)$
- 95% CI for the effect of bill length on mass for Chinstraps?

$$59.11774 \pm (1.967049)(13.60261)$$
$$= 32.36, 85.87$$

```
qt(.975, df = 336)
[1] 1.967049
```

- The estimated effect of bill length on mean mass is statistically significant in Chinstraps $(\mathrm{t} = 4.346, \mathrm{df} = 336, \mathrm{p} < 0.0001)$.
- A one mm increase in bill length in Chinstrap penguins is associated with a 32.4 g to 85.9 g increase in mean weight.

```
59.11774 + c(-1,1)(1.967049)(13.60261)
[1] 32.36074 85.87474
```

Mean response and prediction

- Inference for $\mu_{y|x}$ and $\text{pred}(y \mid x)$ are the same in MLR as SLR!!
- Predicting the mass of a Gentoo with a bill length of 40 mm :

```
predict(peng_interaction_lm,
        newdata = data.frame(bill_length_mm = 40,
                             species = "Gentoo"),
        interval = "prediction")
```

```
       fit      lwr      upr
1 4254.539 3502.729 5006.348
```

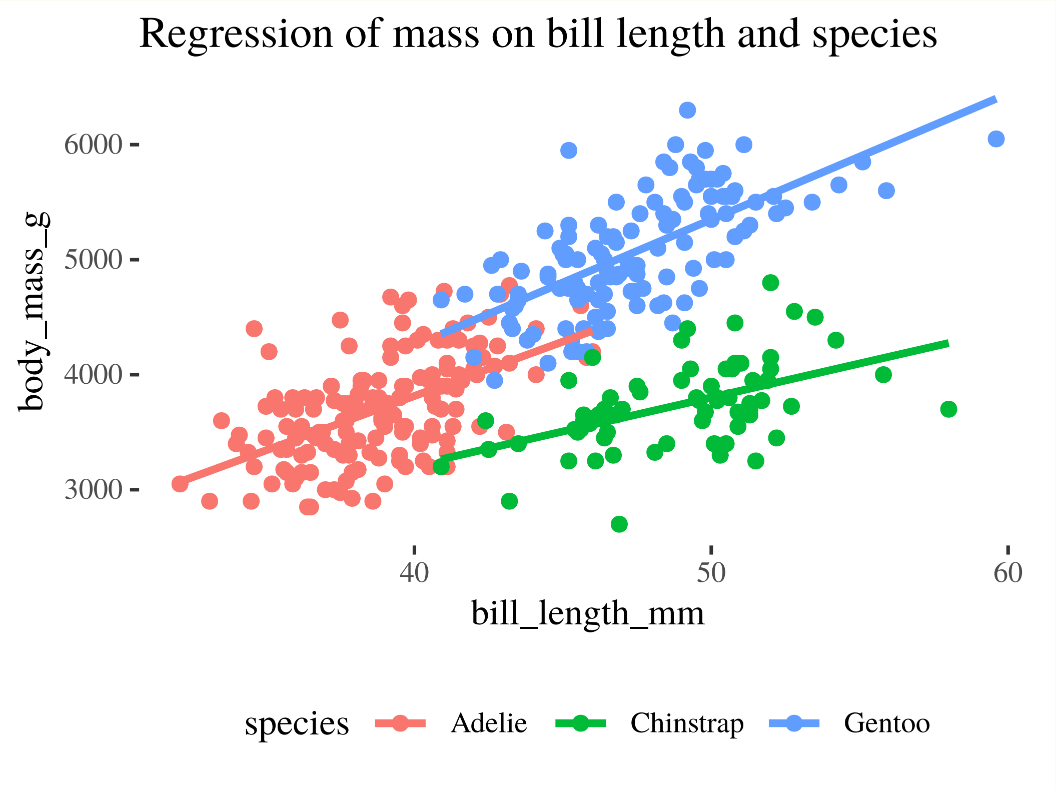
Pengiuns big picture

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	34.883	443.176	0.079	0.937	-836.866	906.632
bill_length_mm	94.500	11.398	8.291	0.000	72.080	116.920
species: Chinstrap	811.260	799.806	1.014	0.311	-761.997	2384.517
species: Gentoo	-158.711	683.191	-0.232	0.816	-1502.582	1185.160
bill_length_mm:speciesChinstrap	-35.382	17.747	-1.994	0.047	-70.291	-0.474
bill_length_mm:speciesGentoo	14.959	15.786	0.948	0.344	-16.093	46.012

- What do the last two rows of t-tests?
- What can we tell or not tell from the last two rows of results?

Pengiuns big picture

Is the interaction of bill length and species needed?



- t-test results for interaction terms tell us *individually* how Chinstrap differs from Adelie β_4 and how Gentoo differs from Adelie β_5
- We can't use t-test results to remove more than one term!
- All t-test results assume *all other terms are in the model!*

Next...

We will test whether **the effect of bill length on mass depends on species**:

$$H_0 : \beta_4 = \beta_5 = 0$$

with a new test done using **ANOVA**!