Model Selection and Inference

Merlise Clyde

September 13, 2017

Last Class

Model for brain weight as a function of body weight

- In the model with both response and predictor log transformed, are dinosaurs outliers?

Last Class

Model for brain weight as a function of body weight

- In the model with both response and predictor log transformed, are dinosaurs outliers?
- should you test each one individually or as a group; if as a group how do you think you would you do this using `lm`?

Last Class

Model for brain weight as a function of body weight

- In the model with both response and predictor log transformed, are dinosaurs outliers?
- should you test each one individually or as a group; if as a group how do you think you would you do this using `lm`?
- do you think your final model is adequate? What else might you change?

Dummy variables

Create an indicator variable for each of the dinosaurs:

```
Animals =
  Animals %>%
    mutate(name = row.names(Animals)) %>%
    mutate(Dino.T = (name ==  "Triceratops")) %>%
    mutate(Dino.D = (name == "Dipliodocus")) %>%
    mutate(Dino.B = (name == "Brachiosaurus")) %>%
    mutate(Dino = (name %in%
                  c("Triceratops",
                    "Brachiosaurus",
                    "Dipliodocus")))
```

uses the `dplyr` package and pipes %>% with `mutate`

New Dataframe

```
##        body brain Dino.T Dino.D Dino.B  Dino             n
## 1      1.35   8.1  FALSE  FALSE  FALSE FALSE Mountain bea
## 2    465.00 423.0  FALSE  FALSE  FALSE FALSE
## 3     36.33 119.5  FALSE  FALSE  FALSE FALSE       Grey w
## 4     27.66 115.0  FALSE  FALSE  FALSE FALSE            G
## 5      1.04   5.5  FALSE  FALSE  FALSE FALSE      Guinea 
## 6  11700.00  50.0  FALSE   TRUE  FALSE  TRUE     Dipliodo
```

Dinosaurs as Outliers

```
brain_out.lm = lm(log(brain) ~ log(body) +
               Dino.T + Dino.B + Dino.D, data=Animals)
kable(summary(brain_out.lm)$coef)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.1504121	0.2006036	10.719710	0.0e+00
log(body)	0.7522607	0.0457186	16.454141	0.0e+00
Dino.TTRUE	-4.7839476	0.7913326	-6.045432	3.6e-06
Dino.BTRUE	-5.6661781	0.8327593	-6.804101	6.0e-07
Dino.DTRUE	-5.2850740	0.7949261	-6.648510	9.0e-07

Anova with Nested Models

- Compare models through Extra-Sum-of Squares

Anova with Nested Models

- Compare models through Extra-Sum-of Squares
  - Each additional predictor reduces the SSE (sum of squares error)

Anova with Nested Models

- Compare models through Extra-Sum-of Squares
  - Each additional predictor reduces the SSE (sum of squares error)
  - Adds to model complexity (more parameters) fewer degrees of freedom for error

Anova with Nested Models

- Compare models through Extra-Sum-of Squares
  - Each additional predictor reduces the SSE (sum of squares error)
  - Adds to model complexity (more parameters) fewer degrees of freedom for error
- Is the addition worth it? Is the decrease "significant"?

$$\frac{\Delta \text{SSE}}{\Delta \text{ df}}$$

Anova with Nested Models

- Compare models through Extra-Sum-of Squares
  - Each additional predictor reduces the SSE (sum of squares error)
  - Adds to model complexity (more parameters) fewer degrees of freedom for error
- Is the addition worth it? Is the decrease "significant"?

$$\frac{\Delta \mathrm{SSE}}{\Delta\ \mathrm{df}}$$

- How big is big enough?

$$F = \frac{\frac{\Delta \mathrm{SSE}}{\Delta\ \mathrm{df}}}{\frac{\mathrm{SSE}_F}{\mathrm{df}_F}} = \frac{\frac{\Delta \mathrm{SSE}}{\Delta\ \mathrm{df}}}{\hat{\sigma}^2} \sim F(\Delta\ \mathrm{df}, n - p)$$

Simultaneous Test: Anova in R

Model:
$\log(brain) = \beta_0 + \log(body)\beta_1 + \text{Dino.T}\beta_2 + \text{Dino.B}\beta_3 + \text{Dino.D}\beta_4 + \epsilon$

Hypothesis Test: $\beta_2 = \beta_3 = \beta_4 = 0$

```
anova(brain_out.lm, brain.lm)
```

```
## Analysis of Variance Table
##
## Model 1: log(brain) ~ log(body) + Dino.T + Dino.B + Dino
## Model 2: log(brain) ~ log(body)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     23 12.117
## 2     26 60.988 -3   -48.871 30.921 3.031e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Other Possible Models

1. all animals follow the same regression

```
brain1.lm = lm(log(brain) ~ log(body),data=Animals) #0
brain2.lm = lm(log(brain) ~ log(body) + Dino,data=Animals)
brain3.lm = lm(log(brain) ~ log(body)*Dino, data=Animals) #
brain4.lm = lm(log(brain) ~ log(body) +
               Dino.T + Dino.B + Dino.D, data=Animals)  #3
```

Other Possible Models

1. all animals follow the same regression
2. the rate of change is the same, but a different intercept for dinosaurs (parallel regression)

```
brain1.lm = lm(log(brain) ~ log(body),data=Animals) #0
brain2.lm = lm(log(brain) ~ log(body) + Dino,data=Animals)
brain3.lm = lm(log(brain) ~ log(body)*Dino, data=Animals) #
brain4.lm = lm(log(brain) ~ log(body) +
               Dino.T + Dino.B + Dino.D, data=Animals)  #3
```

Other Possible Models

1. all animals follow the same regression
2. the rate of change is the same, but a different intercept for dinosaurs (parallel regression)
3. different slopes and intercepts for dinosaurs and other animals

```
brain1.lm = lm(log(brain) ~ log(body),data=Animals) #0
brain2.lm = lm(log(brain) ~ log(body) + Dino,data=Animals)
brain3.lm = lm(log(brain) ~ log(body)*Dino, data=Animals) #
brain4.lm = lm(log(brain) ~ log(body) +
               Dino.T + Dino.B + Dino.D, data=Animals)  #3
```

Other Possible Models

1. all animals follow the same regression
2. the rate of change is the same, but a different intercept for dinosaurs (parallel regression)
3. different slopes and intercepts for dinosaurs and other animals
4. all dinosaurs have a different mean (outliers)

```
brain1.lm = lm(log(brain) ~ log(body),data=Animals) #0
brain2.lm = lm(log(brain) ~ log(body) + Dino,data=Animals)
brain3.lm = lm(log(brain) ~ log(body)*Dino, data=Animals) #
brain4.lm = lm(log(brain) ~ log(body) +
               Dino.T + Dino.B + Dino.D, data=Animals)  #3
```

Sequential Sum of Squares

```r
anova(brain1.lm, brain2.lm, brain3.lm, brain4.lm)
```

```
## Analysis of Variance Table
##
## Model 1: log(brain) ~ log(body)
## Model 2: log(brain) ~ log(body) + Dino
## Model 3: log(brain) ~ log(body) * Dino
## Model 4: log(brain) ~ log(body) + Dino.T + Dino.B + Dino
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     26 60.988
## 2     25 12.505  1    48.483 92.0248 1.665e-09 ***
## 3     24 12.212  1     0.294  0.5578    0.4627
## 4     23 12.117  1     0.094  0.1788    0.6763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Model Selection

- Fail to reject Model 3 in favor of Model 4

Model Selection

- Fail to reject Model 3 in favor of Model 4
- Fail to reject Model 2 in favor of Model 3

Model Selection

- Fail to reject Model 3 in favor of Model 4
- Fail to reject Model 2 in favor of Model 3
- Reject Model 1 in favor of Model 2

Model Selection

- Fail to reject Model 3 in favor of Model 4
- Fail to reject Model 2 in favor of Model 3
- Reject Model 1 in favor of Model 2
  - Same slope for `log(body)` for all animals, but different intercepts

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.16	0.19	11.09	0.00
log(body)	0.75	0.04	16.90	0.00
DinoTRUE	-5.22	0.53	-9.84	0.00

Distribution of Coefficients

- Joint Distribution under normality

$$\hat{\boldsymbol{\beta}} \mid \sigma^2 \sim \mathsf{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

Distribution of Coefficients

- Joint Distribution under normality

$$\hat{\boldsymbol{\beta}} \mid \sigma^2 \sim \mathsf{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

- Distribution of SSE

$$\mathsf{SSE} \sim \chi^2(n-p)$$

Distribution of Coefficients

- Joint Distribution under normality

$$\hat{\boldsymbol{\beta}} \mid \sigma^2 \sim \mathsf{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

- Distribution of SSE

$$\mathsf{SSE} \sim \chi^2(n-p)$$

- Marginal distribution

$$\frac{\hat{\beta}_j - \beta_j}{\mathsf{SE}(\hat{\beta}_j)} \sim \mathsf{St}(n-p)$$

$$\mathsf{SE}(\hat{\beta}_j) = \hat{\sigma}\sqrt{[\mathbf{X}^T\mathbf{X}]^{-1}]_{jj}]}$$

Confidence Intervals

$(1-\alpha/2)100\%$ Confidence interval for β_j

$$\hat{\beta}_j \pm t_{n-p,\alpha/2}\mathrm{SE}(\hat{\beta}_j)$$

```
kable(confint(brain2.lm))
```

	2.5 %	97.5 %
(Intercept)	1.760198	2.5630848
log(body)	0.657346	0.8397591
DinoTRUE	-6.311226	-4.1274760

Converting to Original Units

- Model after exponentiating

$$
\begin{aligned}
\widehat{brain} &= e^{\hat{\beta}_0 + \log(body)\hat{\beta}_1 + \text{Dino}\hat{\beta}_2} \\
&= e^{\hat{\beta}_0} e^{\log(body)\hat{\beta}_1} e^{\text{Dino}\hat{\beta}_2} \\
&= e^{\hat{\beta}_0} body^{\hat{\beta}_1} e^{\text{Dino}\hat{\beta}_2}
\end{aligned}
$$

Converting to Original Units

- Model after exponentiating

$$\begin{aligned}\widehat{brain} &= e^{\hat{\beta}_0 + \log(body)\hat{\beta}_1 + \text{Dino}\hat{\beta}_2} \\ &= e^{\hat{\beta}_0} e^{\log(body)\hat{\beta}_1} e^{\text{Dino}\hat{\beta}_2} \\ &= e^{\hat{\beta}_0} body^{\hat{\beta}_1} e^{\text{Dino}\hat{\beta}_2}\end{aligned}$$

- 10% increase in body weight implies a

$$\begin{aligned}\widehat{brain}_{1.10} &= e^{\hat{\beta}_0} (1.10 * body)^{\hat{\beta}_1} e^{\text{Dino}\hat{\beta}_2} \\ &= 1.10^{\hat{\beta}_1} e^{\hat{\beta}_0} body^{\hat{\beta}_1} e^{\text{Dino}\hat{\beta}_2}\end{aligned}$$

Converting to Original Units

- Model after exponentiating

$$
\begin{aligned}
\widehat{brain} &= e^{\hat{\beta}_0 + \log(body)\hat{\beta}_1 + \text{Dino}\hat{\beta}_2} \\
&= e^{\hat{\beta}_0} e^{\log(body)\hat{\beta}_1} e^{\text{Dino}\hat{\beta}_2} \\
&= e^{\hat{\beta}_0} body^{\hat{\beta}_1} e^{\text{Dino}\hat{\beta}_2}
\end{aligned}
$$

- 10% increase in body weight implies a

$$
\begin{aligned}
\widehat{brain}_{1.10} &= e^{\hat{\beta}_0} (1.10 * body)^{\hat{\beta}_1} e^{\text{Dino}\hat{\beta}_2} \\
&= 1.10^{\hat{\beta}_1} e^{\hat{\beta}_0} body^{\hat{\beta}_1} e^{\text{Dino}\hat{\beta}_2}
\end{aligned}
$$

- $1.1^{\hat{\beta}_1} = 1.074$ or a 7.4% increase in brain weight

95% Confidence interval

To obtain a 95% confidence interval, $(1.10^{CI} - 1) * 100$

	2.5 %	97.5 %
body	6.465603	8.332779

Interpretation of Intercept

- Evalutate model with predictors $= 0$

$$\widehat{\log(brain)} = \hat{\beta}_0 + \log(body)\hat{\beta}_1 + \text{Dino}\hat{\beta}_2$$

Interpretation of Intercept

- Evalutate model with predictors $= 0$

$$\widehat{\log(brain)} = \hat{\beta}_0 + \log(body)\hat{\beta}_1 + \text{Dino}\hat{\beta}_2$$

- For a non-dinosaur, if `log(body) = 0` (body weight $= 1$ kilogram), we expect that brain weight will be 2.16 log(grams) ???

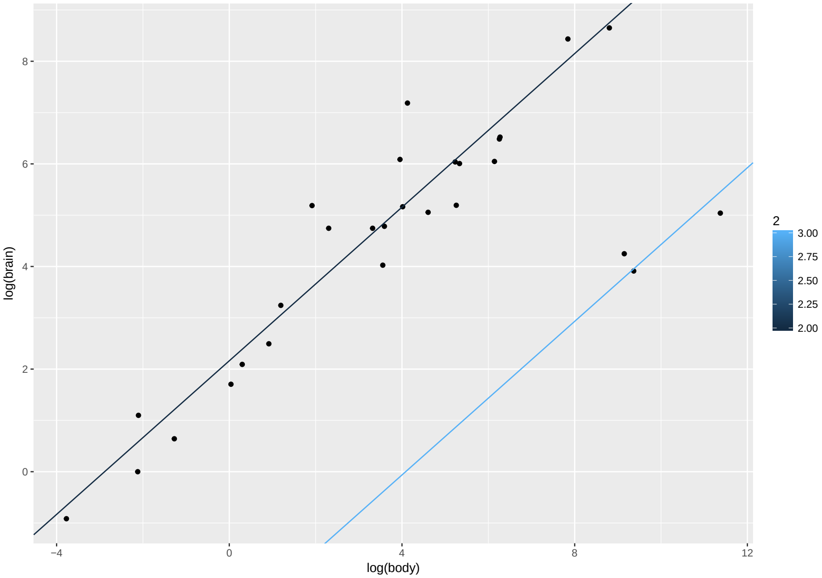
Interpretation of Intercept

- Evalutate model with predictors $= 0$

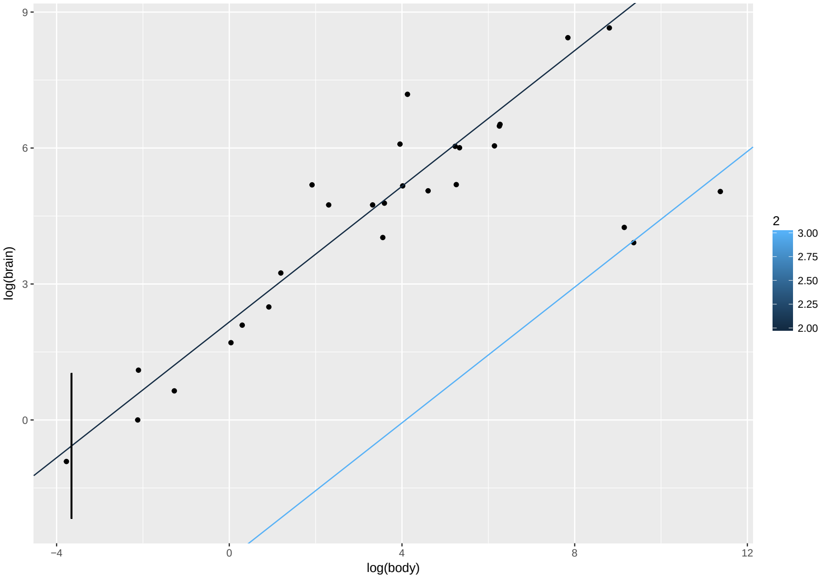
$$\widehat{\log(brain)} = \hat{\beta}_0 + \log(body)\hat{\beta}_1 + \text{Dino}\hat{\beta}_2$$

- For a non-dinosaur, if `log(body) = 0` (body weight $= 1$ kilogram), we expect that brain weight will be 2.16 log(grams) ???
- Exponentiate: predicted brain weight for non-dinosaur with a 1 kg body weight is
$$e^{\hat{\beta}_0} = 8.69 \text{ grams}$$

Plot of Fitted Values



Predictions for 259 gram cockatoo



Predictions in original units

- 95% Confidence Interval for $f(x)$

```
newdata = data.frame(body=.0259, Dino=FALSE)
fit = predict(brain2.lm, newdata=newdata,
              interval="confidence", se=T)
```

```
pred = predict(brain2.lm, newdata=newdata,
               interval="predict", se=T)
```

Predictions in original units

- 95% Confidence Interval for $f(x)$

```
newdata = data.frame(body=.0259, Dino=FALSE)
fit = predict(brain2.lm, newdata=newdata,
              interval="confidence", se=T)
```

- 95% Prediction Interval for Brain Weight

```
pred = predict(brain2.lm, newdata=newdata,
               interval="predict", se=T)
```

CI/Predictions in original units for body=259 g

- 95% Confidence Interval for $f(x)$

```
exp(fit$fit)
```

```
##         fit       lwr      upr
## 1 0.5637161 0.2868832 1.107684
```

```
exp(pred$fit)
```

```
##         fit       lwr     upr
## 1 0.5637161 0.1131737 2.80786
```

CI/Predictions in original units for body=259 g

- 95% Confidence Interval for $f(x)$

```
exp(fit$fit)
```

```
##         fit       lwr      upr
## 1 0.5637161 0.2868832 1.107684
```

- 95% Prediction Interval for Brain Weight

```
exp(pred$fit)
```

```
##         fit       lwr     upr
## 1 0.5637161 0.1131737 2.80786
```

CI/Predictions in original units for body=259 g

- 95% Confidence Interval for $f(x)$

```r
exp(fit$fit)
```

```
##         fit       lwr      upr
## 1 0.5637161 0.2868832 1.107684
```

- 95% Prediction Interval for Brain Weight

```r
exp(pred$fit)
```

```
##         fit       lwr     upr
## 1 0.5637161 0.1131737 2.80786
```

- 95% confident that the brain weight will be between 0.11 and 2.81 grams

Summary

- Linear predictors may be based on functions of other predictors (dummy variables, interactions, non-linear terms)

Summary

- Linear predictors may be based on functions of other predictors (dummy variables, interactions, non-linear terms)
- need to change back to original units

Summary

- Linear predictors may be based on functions of other predictors (dummy variables, interactions, non-linear terms)
- need to change back to original units
- log transform useful for non-negative responses (ensures predictions are non-negative)

Summary

- Linear predictors may be based on functions of other predictors (dummy variables, interactions, non-linear terms)
- need to change back to original units
- log transform useful for non-negative responses (ensures predictions are non-negative)
- Be careful of units of data

Summary

- Linear predictors may be based on functions of other predictors (dummy variables, interactions, non-linear terms)
- need to change back to original units
- log transform useful for non-negative responses (ensures predictions are non-negative)
- Be careful of units of data
  - plots should show units

Summary

- Linear predictors may be based on functions of other predictors (dummy variables, interactions, non-linear terms)
- need to change back to original units
- log transform useful for non-negative responses (ensures predictions are non-negative)
- Be careful of units of data
  - plots should show units
  - summary statements should include units

Summary

- Linear predictors may be based on functions of other predictors (dummy variables, interactions, non-linear terms)
- need to change back to original units
- log transform useful for non-negative responses (ensures predictions are non-negative)
- Be careful of units of data
  - plots should show units
  - summary statements should include units
- Goodness of fit measure: R^2 and Adjusted R^2 depend on scale
  R^2 is percent variation in "Y" that is explained by the model

$$R^2 = 1 - SSE/SST$$

where SST $= \sum_i (Y_i - \bar{Y})^2$