# Atividade 7

**Verificando as suposições do modelo**

Paulo Ricardo Seganfredo Campana

29 de setembro de 2023

Questão 1.

Por razão de segurança, um estudo de ursos envolveu a coleção de várias medidas, que foram feitas depois que os ursos foram anestesiados. É relativamente fácil usar uma fita métrica para encontrar algumas medidas, como tamanho da cabeça, tamanho do tórax e comprimento do urso, mas é difícil pesar o urso. Em vez de pesar o urso, podemos predizer seu peso com com base nestas outras medidas? Considere a seguinte tabela de dados:

| Peso | 80 | 344 | 416 | 348 | 262 | 360 | 332 | 34 | 140 | 180 | 105 | 166 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tmn. da cabeça | 11 | 16.5 | 15.5 | 17 | 15 | 13.5 | 16 | 9 | 12.5 | 14 | 11.5 | 13 |
| Comprimento | 53 | 67.5 | 72 | 72 | 73.5 | 68.5 | 73 | 37 | 63 | 67 | 52 | 59 |
| Tmn. do tórax | 26 | 45 | 54 | 49 | 41 | 49 | 44 | 19 | 32 | 37 | 29 | 33 |

a) Estime um modelo de regressão linear múltiplo que explique os pesos dos ursos em função das variáveis: tamanho da cabeça, tamanho do tórax e comprimento do urso.

```
fit1 <- lm(y ~ x1 + x2 + x3, data)
coefficients(fit1)
## (Intercept)          x1          x2          x3
## -203.660398    9.299429   -2.748157   12.582746
```

O modelo estimado é $\hat{y} = -203.66 + 9.30x_1 - 2.75x_2 + 12.58x_3$.

b) Através do teste $F$, você acha que o modelo adotado é razoável? Justifique sua resposta.

Sim, conjuntamente o modelo é significativo com p-valor de $4.41 \times 10^{-7}$ no teste $F$, porém alguns regressores parecem não estarem contribuindo muito para o modelo.

Tabela 2: Teste $F$ para o modelo de regressão linear múltipla

| Variável | gl | SS | MS | Estatística | p-valor |
|---|---|---|---|---|---|
| Tamanho da cabeça | 1 | 140985 | 140985 | 302.11 | $1.22 \times 10^{-7}$ |
| Comprimento | 1 | 2379 | 2379 | 5.10 | 0.0539 |
| Tamanho do tórax | 1 | 33980 | 33980 | 72.81 | 0.0000274 |
| (Regressão) | 3 | 177344 | 59115 | 126.67 | $4.41 \times 10^{-7}$ |
| (Resíduos) | 8 | 3733 | 467 | | |
| (Total) | 11 | 181077 | 16462 | | |

c) O $R^2$ e $R_a^2$ sugerem que o modelo proposto explica razoavelmente os dados?

```
summary(fit1)$r.squared
## [1] 0.9793824
summary(fit1)$adj.r.squared
## [1] 0.9716509
```

Sim, temos um coeficiente de determinação ajustado muito alto, mesmo com poucas observações e 3 variáveis.

d) Analisando os testes $t$ para cada coeficiente da regressão, quais variáveis deverão ser excluídas?

Tabela 3: Teste $t$ para as variáveis do modelo de regressão linear múltipla

| Variável | Estimativa | Erro padrão | Estatística | p-valor |
|---|---|---|---|---|
| (Intercepto) | -203.66 | 46.30 | -4.40 | 0.00229 |
| Tamanho da cabeça | 9.30 | 7.15 | 1.30 | 0.23 |
| Comprimento | -2.75 | 1.61 | -1.70 | 0.127 |
| Tamanho do tórax | 12.58 | 1.48 | 8.53 | 0.0000274 |

Os regressores Tamanho da cabeça e Comprimento não são significantes segundo o teste $t$ então devem ser retirados do modelo.

e) O pesquisador que lidera este estudo não entende porque nem todas as variáveis foram significativas. Verifique a hipótese de multicolinearidade e, com base nesses resultados, justifique a exclusão dessas variáveis.

```
car::vif(fit1)
##       x1       x2       x3
## 7.153588 7.529552 5.776141
```

Temos Fatores de Inflação da Variância altos (maiores que 5) para todas as variáveis, isso indica multicolinearidade na regressão, iremos então retirar as variáveis Tamanho da cabeça e Comprimento.

f) Depois dessa análise e considerando que não é possível obter novos dados, ajuste um modelo de regressão que explique o peso através do tamanho do tórax. Considerando esta regressão faça as seguintes análises:

```
fit2 <- lm(y ~ x3, data)
```

g) Verifique a hipótese de normalidade. Comente os resultados.

Tabela 4: Resultados dos testes de normalidade para os resíduos studentizados

| Teste de normalidade | Estatística | p-valor |
|---|---|---|
| Lilliefors | $D = 0.262$ | 0.0221 |
| Shapiro-Wilk | $W = 0.898$ | 0.1517 |
| Jarque-Bera | $JB = 0.504$ | 0.7771 |

Apenas o teste de normalidade de Lilliefors rejeita a normalidade dos resíduos, provavelmente devido ao pequeno número de observações.

h) Verifique a hipótese de linearidade. Comente os resultados.

```
data |>
    mutate(predict = predict(fit2)) |>
    ggplot(aes(x = y, y = predict)) +
    geom_point(color = "#4582ec") +
    geom_abline(intercept = 0, slope = 1, alpha = 0.25) +
    theme_bw()
```

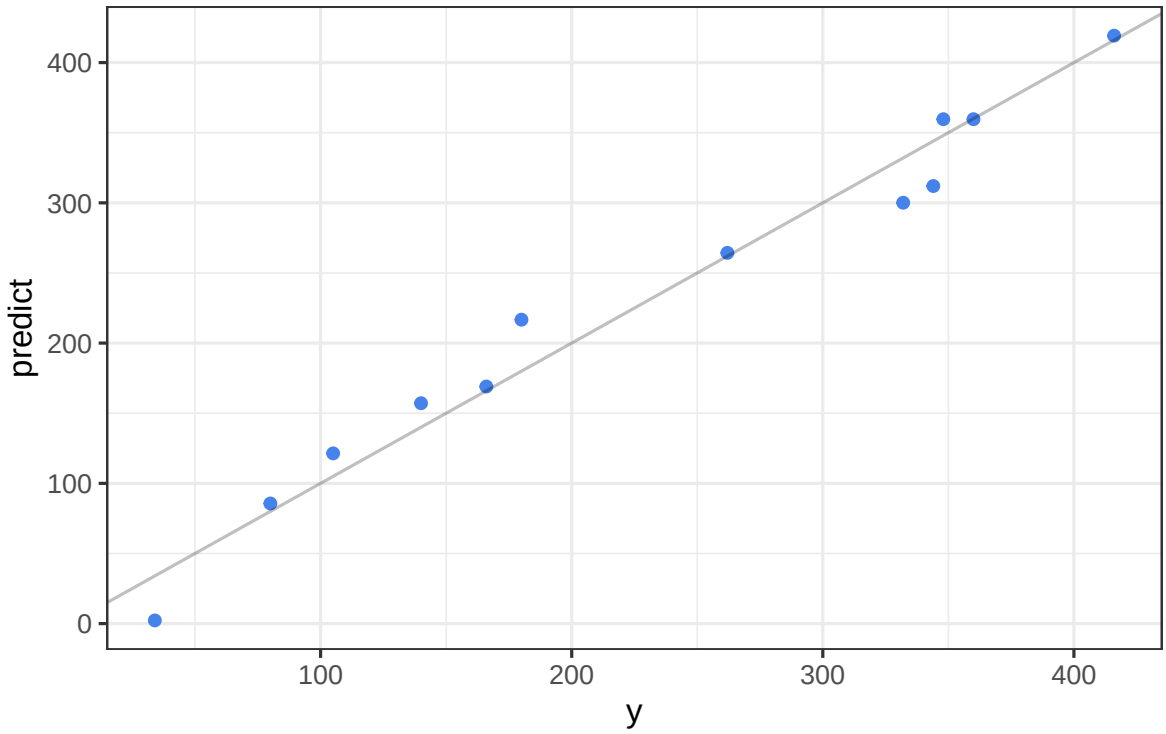


Tabela 5: Resultados dos testes de linearidade

| Teste de linearidade | Estatística | p-valor |
|---|---|---|
| RESET (y) | RESET = 3.661 | 0.0743 |
| RESET ($x_3$) | RESET = 3.661 | 0.0743 |
| Rain test | Rain = 1.094 | 0.4876 |

Pelo gráfico, vemos um comportamento bastante linear entre $y$ e $\hat{y}$ além de que os testes não rejeitam a hipótese de linearidade.

i) Verifique a hipótese de autocorrelação. Comente os resultados.

```
lmtest::dwtest(fit2)$p.value
## [1] 0.1043301
```

O teste para autocorrelação de Durbin-Watson indica que não há autocorrelação no modelo.

j) Verifique a hipótese de homocedasticidade. Comente os resultados.

Tabela 6: Resultados dos testes para heterocedasticidade

| Teste de heterocedasticidade | Estatística | p-valor |
|---|---|---|
| Goldfeld-Quandt | GQ = 3.244 | 0.140 |
| Breusch-Pagan | BP = 0.264 | 0.607 |
| Koenker | BP = 0.405 | 0.524 |

Todos os testes indicam que o modelo é homocedástico a um nível de significância de 5%.

k) Através dos itens anteriores e do $R^2$, teste $F$ e testes $t$, você acha que o modelo adotado é razoável? Justifique sua resposta.

```
summary(fit2)$adj.r.squared
## [1] 0.968706
```

Sim, o modelo é razoável pois continuamos com um $R^2$ muito alto e o novo modelo é muito mais significativo com apenas um regressor.