# Modelos de Regressão e Previsão

Lista 2 - Gabarito

Prof. Carlos Trucíos
carlos.trucios@facc.ufrj.br
ctruciosm.github.io

- Caso se indique o contrário, considere sempre um nível de significância de $\alpha = 0.05$
- Os comandos *summary()*, *confint()*, *anova()* e *predict()* lhe ajudarão a responder as perguntas
- Verifique o pacote *wooldridge* está instalado e carregado

```
install.packages("wooldridge")
library(wooldridge)
```

**Questão 1**

No seguinte modelo:

```
modelo = lm(log(bwght) ~ npvis + I(npvis^2), data = bwght2)
```

- As variáveis são estatísticamente significativas? Sim (p-valor $< 0.05$)

```
summary(modelo)
```

```
##
## Call:
## lm(formula = log(bwght) ~ npvis + I(npvis^2), data = bwght2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15564 -0.08375  0.02241  0.11417  0.45529
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.9578826  0.0273125 291.364  < 2e-16 ***
## npvis        0.0189167  0.0036806   5.140 3.06e-07 ***
## I(npvis^2)  -0.0004288  0.0001200  -3.573 0.000362 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2031 on 1761 degrees of freedom
##   (68 observations deleted due to missingness)
## Multiple R-squared:  0.02125,    Adjusted R-squared:  0.02014
## F-statistic: 19.12 on 2 and 1761 DF,  p-value: 6.097e-09
```

- Interprete os betas

```
modelo
```

```
##
## Call:
## lm(formula = log(bwght) ~ npvis + I(npvis^2), data = bwght2)
```

```
##
## Coefficients:
## (Intercept)        npvis   I(npvis^2)
##   7.9578826    0.0189167   -0.0004288
```

O modelo é da forma:

$$\widehat{\log(bwght)} = 7.9578826 + 0.0189167\text{npvis} - 0.0004288\text{npvis}^2$$

Então,

$$\frac{\Delta\widehat{\log(bwght)}}{\Delta\text{npvis}} \approx 0.0189167 - \underbrace{2 \times 0.0004288}_{0.0008576}\text{npvis}$$

$$\frac{\%\Delta bwght}{\Delta\text{npvis}} \approx \frac{100\Delta\widehat{\log(bwght)}}{\Delta\text{npvis}} \approx 1.89167 - 0.08576\text{npvis}$$

O número de visitas prenatais (npvis) tem um efeito positivo no peso da criança ao nascer, embora o efeito não é constante: - $1.89167 - 0.08576(1) = 1.80591$ - $1.89167 - 0.08576(2) = 1.72015$ - $1.89167 - 0.08576(3) = 1.63439$ - $1.89167 - 0.08576(9) = 1.11983$

veja que o ponto de infleção é $1.89167/0.08576 = 22.05772$

```
prop.table(table(bwght2$npvis>22))
```

```
##
##      FALSE       TRUE
## 0.98922902 0.01077098
```

- Calcula intervalos de confianca 90% para os betas

```
confint(modelo, level = 0.90)
```

```
##                        5 %          95 %
## (Intercept)  7.9129338434  8.0028313888
## npvis        0.0128595115  0.0249739471
## I(npvis^2)  -0.0006262304 -0.0002312825
```

**Questão 2**

No seguinte modelo:

```
modelo = lm(log(wage)~educ+exper+tenure+nonwhite+female+
              married+numdep, data = wage1)
```

- Qual percentagem da variabilidade de $\log(wage)$ é explicada pelo modelo?

```
summary(modelo)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + exper + tenure + nonwhite + female +
##     married + numdep, data = wage1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.87251 -0.27290 -0.03787  0.25332  1.23647
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  0.4898493  0.1096290   4.468 9.69e-06 ***
## educ         0.0839250  0.0072943  11.506  < 2e-16 ***
## exper        0.0031410  0.0017177   1.829  0.06804 .
## tenure       0.0168677  0.0029619   5.695 2.07e-08 ***
## nonwhite    -0.0026234  0.0598846  -0.044  0.96507
## female      -0.2856052  0.0373891  -7.639 1.07e-13 ***
## married      0.1254258  0.0413853   3.031  0.00256 **
## numdep       0.0003184  0.0152552   0.021  0.98336
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4133 on 518 degrees of freedom
## Multiple R-squared:  0.4036, Adjusted R-squared:  0.3956
## F-statistic: 50.08 on 7 and 518 DF,  p-value: < 2.2e-16
```

Aproximadamente 40% da variabilidade de $\log(wage)$ é explicada pelo modelo[1]

- Quais variáveis são estatísticamente significativas? Aquelas com p-valor $< 0.05$,
  - educ
  - tenure
  - female
  - married
- Teste a hipótese: $H_0 : \beta_{nonwhite} = 0, \beta_{numdep} = 0$ vs. $H_1 : H_0$ não é verdadeira . Podemos rejeitar $H_0$?

```
modeloi = modelo
modelor = lm(log(wage)~educ+exper+tenure+female+married, data = wage1)
anova(modelor,modeloi)
```

```
## Analysis of Variance Table
##
## Model 1: log(wage) ~ educ + exper + tenure + female + married
## Model 2: log(wage) ~ educ + exper + tenure + nonwhite + female + married +
##     numdep
##   Res.Df    RSS Df  Sum of Sq      F Pr(>F)
## 1    520 88.462
## 2    518 88.462  2 0.00038039 0.0011 0.9989
```

A um nível de significância de 0.05 **não rejeitamos** $H_0$

- No modelo de regressão original, inclua o termo de interação nonwhite*female*married. Alguma interação é estatisticamente significativa?

```
modelo = lm(log(wage)~educ+exper+tenure+nonwhite+female+
              married+numdep + nonwhite*female*married, data = wage1)
summary(modelo)
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + exper + tenure + nonwhite + female +
##     married + numdep + nonwhite * female * married, data = wage1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.97059 -0.24239 -0.03676  0.23630  1.18320
##
## Coefficients:
```

[1] Geralmente preferimos o $R^2_{adjustado}$

```
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.398370   0.112296   3.547 0.000425 ***
## educ                      0.082891   0.007252  11.430  < 2e-16 ***
## exper                     0.003089   0.001704   1.812 0.070516 .
## tenure                    0.015647   0.002944   5.314 1.60e-07 ***
## nonwhite                  0.017737   0.144552   0.123 0.902391
## female                   -0.082264   0.061782  -1.332 0.183611
## married                   0.297547   0.060435   4.923 1.15e-06 ***
## numdep                   -0.006711   0.015196  -0.442 0.658924
## nonwhite:female          -0.084353   0.180315  -0.468 0.640119
## nonwhite:married          0.004450   0.173353   0.026 0.979529
## female:married           -0.325599   0.079073  -4.118 4.46e-05 ***
## nonwhite:female:married   0.012225   0.252957   0.048 0.961472
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4074 on 514 degrees of freedom
## Multiple R-squared:  0.4247, Adjusted R-squared:  0.4124
## F-statistic:  34.5 on 11 and 514 DF,  p-value: < 2.2e-16
```

A um nível de significância de 0.05, a interação female×married é estatísticamente significativa

**Questão 3**

Sejam os seguintes modelos:

```
modelo1 = lm(log(wage)~educ+exper+tenure+nonwhite+female+married+numdep, data = wage1)
modelo2 = lm(log(wage)~educ+exper+tenure+female+married + smsa + construc +
               ndurman + trade + services + profserv + profocc + servocc, data = wage1)
modelo3 = lm(log(wage)~educ+exper+ I(exper^2) + tenure+female+married +
               smsa + trade + services + profocc + servocc + female*married, data = wage1)
```

- Se seu interesse é explicar a variabilidade de $\log(wage)$, qual modelo escolheria? (porquê?) Vamos escolher o modelo com maior $R^2_{adj}$

```
summary(modelo1)$adj.r.squared
```

```
## [1] 0.3955544
```

```
summary(modelo2)$adj.r.squared
```

```
## [1] 0.5021672
```

```
summary(modelo3)$adj.r.squared
```

```
## [1] 0.5366229
```

Preferimos o modelo3, pois a percentagem de variabilidade de $\log(wage)$ que é explicada pelo modelo é maior.

- Se seu interesse é explicar a variabilidade de $wage$, qual modelo escolheria? (porquê?) Vamos utilizar a transformação vista na aula para obter $\widehat{wage}$ e comparar os modelos utilizando $R^2$, ge

$$\widehat{wage} = \hat{\alpha}\, \underbrace{e^{\widehat{\log(wage)}}}_{m}$$

onde $\hat{\alpha} = n^{-1}\sum_{i=1}^{n}\exp(\hat{u}_i)$ ou $\hat{\alpha} = \left(\sum_{i=1}^{n}\hat{m}_i y_i\right)\Big/\left(\sum_{i=1}^{n}\hat{m}_i^2\right)$ onde $\hat{m}_i = \exp(\widehat{\log(y_i)})$

```
uhat1 = residuals(modelo1)
uhat2 = residuals(modelo2)
```

```
uhat3 = residuals(modelo3)

alpha1 = mean(exp(uhat1))
alpha2 = mean(exp(uhat2))
alpha3 = mean(exp(uhat3))

m1 = exp(fitted(modelo1))
m2 = exp(fitted(modelo2))
m3 = exp(fitted(modelo3))

yhat1 = alpha1*m1
yhat2 = alpha2*m2
yhat3 = alpha3*m3

y = wage1$wage
1- sum((y-yhat1)^2)/sum((y-mean(y))^2)
```

```
## [1] 0.3963844
```

```
1- sum((y-yhat2)^2)/sum((y-mean(y))^2)
```

```
## [1] 0.4901768
```

```
1- sum((y-yhat3)^2)/sum((y-mean(y))^2)
```

```
## [1] 0.5316782
```

Escolhemos o modelo3

Mas professor, e se nos modelos acima utilizarmos *wage* em lugar de log(*wage*)?

```
modelo1 = lm(wage~educ+exper+tenure+nonwhite+female+married+numdep, data = wage1)
modelo2 = lm(wage~educ+exper+tenure+female+married + smsa + construc +
               ndurman + trade + services + profserv + profocc + servocc, data = wage1)
modelo3 = lm(wage~educ+exper+ I(exper^2) + tenure+female+married +
               smsa + trade + services + profocc + servocc + female*married, data = wage1)
summary(modelo1)$adj.r.squared
```

```
## [1] 0.3619087
```

```
summary(modelo2)$adj.r.squared
```

```
## [1] 0.4396346
```

```
summary(modelo3)$adj.r.squared
```

```
## [1] 0.4773257
```

Escolhemos o **novo modelo3**, mas note que a percentagem de variabilidade explicada pelo **novo modelo3** é menor do que com o modelo3 que utiliza log(*wage*) como variável dependente.

**Questão 4**

No seguinte modelo:

```
modelo = lm(bwght~. -moth -foth, data = bwght2)
```

- Qual percentagem da variabilidade de *bwght* é explicada pelo modelo?

```
summary(modelo)
```

```
##
## Call:
## lm(formula = bwght ~ . - moth - foth, data = bwght2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -773.08  -37.49  -18.85   14.42 1782.75
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -2.300e+04  1.303e+02 -176.538  < 2e-16 ***
## mage         2.815e+00  4.238e+00    0.664  0.50664
## meduc        1.283e+00  1.312e+00    0.978  0.32839
## monpre      -2.002e+00  1.988e+00   -1.007  0.31419
## npvis       -8.227e-01  1.779e+00   -0.463  0.64378
## fage        -1.653e-01  5.223e-01   -0.316  0.75171
## feduc       -1.782e-01  1.186e+00   -0.150  0.88053
## omaps       -2.707e+00  2.361e+00   -1.147  0.25176
## fmaps       -1.571e+01  5.430e+00   -2.893  0.00386 **
## cigs        -1.300e+00  5.433e-01   -2.393  0.01685 *
## drink       -4.289e-02  7.252e+00   -0.006  0.99528
## lbw          4.762e+02  2.428e+01   19.614  < 2e-16 ***
## vlbw         9.646e+02  3.888e+01   24.812  < 2e-16 ***
## male         4.412e+00  4.279e+00    1.031  0.30274
## mwhte       -8.389e+00  1.928e+01   -0.435  0.66354
## mblck       -9.972e+00  2.952e+01   -0.338  0.73556
## fwhte        2.166e+01  2.090e+01    1.036  0.30021
## fblck        7.023e+00  3.037e+01    0.231  0.81715
## lbwght       3.265e+03  1.386e+01  235.507  < 2e-16 ***
## magesq      -4.155e-02  6.997e-02   -0.594  0.55270
## npvissq      4.198e-02  5.534e-02    0.759  0.44816
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.82 on 1591 degrees of freedom
##   (220 observations deleted due to missingness)
## Multiple R-squared:  0.9775, Adjusted R-squared:  0.9772
## F-statistic:  3461 on 20 and 1591 DF,  p-value: < 2.2e-16
```

97.2% da variabilidade de *bwght* é explicada pelo modelo

- Observa alguma coisa errana na modelagem? Sim, o modelo ajustou-se quase perfeitamente aos dados
- Reestime o modelo excluindo também a variavel *lbwght*

```
modelo = lm(bwght~. -moth -foth -lbwght, data = bwght2)
summary(modelo)
```

```
##
## Call:
## lm(formula = bwght ~ . - moth - foth - lbwght, data = bwght2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1444.44  -335.17   -14.21   330.95  1757.02
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.435e+03  4.718e+02   3.042  0.00239 **
## mage         5.198e+01  2.534e+01   2.051  0.04043 *
## meduc        2.413e+00  7.856e+00   0.307  0.75871
## monpre       1.841e+01  1.189e+01   1.548  0.12179
## npvis        1.004e+01  1.064e+01   0.943  0.34566
## fage         5.914e+00  3.123e+00   1.894  0.05845 .
## feduc        4.519e+00  7.097e+00   0.637  0.52440
## omaps        1.271e-01  1.413e+01   0.009  0.99282
## fmaps        6.475e+01  3.244e+01   1.996  0.04613 *
## cigs        -7.825e+00  3.249e+00  -2.409  0.01612 *
## drink       -2.749e+01  4.341e+01  -0.633  0.52667
## lbw         -1.611e+03  1.353e+02 -11.903  < 2e-16 ***
## vlbw        -4.502e+02  2.299e+02  -1.958  0.05040 .
## male         6.511e+01  2.557e+01   2.546  0.01098 *
## mwhte       -1.429e+02  1.154e+02  -1.238  0.21573
## mblck       -3.389e+02  1.765e+02  -1.920  0.05506 .
## fwhte        4.083e+02  1.247e+02   3.273  0.00109 **
## fblck        5.849e+02  1.812e+02   3.228  0.00127 **
## magesq      -9.170e-01  4.183e-01  -2.192  0.02850 *
## npvissq     -8.353e-02  3.312e-01  -0.252  0.80094
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 507.8 on 1592 degrees of freedom
##   (220 observations deleted due to missingness)
## Multiple R-squared:  0.1942, Adjusted R-squared:  0.1846
## F-statistic:  20.2 on 19 and 1592 DF,  p-value: < 2.2e-16
```

- Qual percentagem da variabilidade de *bwght* é explicada pelo modelo? 18.46%
- Quais variáveis são estatísticamente significativas?
  - mage
  - fmaps
  - cigs
  - lbw
  - male
  - fwhte
  - fblck
  - magesq
- Explique essa grande diferença na percentagem de variabilidade explicada pelos modelos.

No primeiro modelo, a variavel $lbwght = \log(bwght)$ é incluida como variavel explicativa para modelar $bwght$. Ou seja para estimar $y$, precisamos de $\log(y)$ o que não faz sentido nenhum, esse é um exemplo claro de **overfitting**

**Questão 5**

No seguinte modelo:

```
modelo = lm(lavgsal ~ bs, data = benefits)
```

- Seja $H_0 : \beta_{bs} = 0$ vs $H_1 : \beta_{bs} \neq 0$, podemos rejeitar $H_0$? Sim, rejeitamos $H_0$ (com um nível de significância de 0.05)

```
summary(modelo)
```

```
##
```

```
## Call:
## lm(formula = lavgsal ~ bs, data = benefits)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.35039 -0.14368  0.00689  0.14759  0.74891
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.64757    0.05724  186.02  < 2e-16 ***
## bs          -0.50346    0.16615   -3.03  0.00248 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2324 on 1846 degrees of freedom
## Multiple R-squared:  0.004949,   Adjusted R-squared:  0.00441
## F-statistic: 9.182 on 1 and 1846 DF,  p-value: 0.002478
```

- Teste a hipótese $H_0 : \beta_{bs} = -1$ ($H_0 : \beta_{bs} \leq -1$) vs $H_1 : \beta_{bs} > -1$, podemos rejeitar $H_0$?

```
# Como a saida padrão testa H0: beta = 0 vs H1: beta != 0
# não podemos utilizas nem a estatistica t nem  p-valor
(-0.50346+1)/0.16615
```

```
## [1] 2.988504
```

```
dim(benefits) #n grande, podemos aproximar por uma Normal
```

```
## [1] 1848   18
```

```
qnorm(0.95)
```

```
## [1] 1.644854
```

2.988504 é maior que 1.644854 ("cai na área cinza"), rejeitamos $H_0$ (com um nível de significância de 5%)

- Estime o modelo

$$lavgsal = \beta_0 + \beta_1 bs + \beta_2 lenroll + \beta_3 lstaff + \beta_4 lunch + u$$

```
modelo = lm(lavgsal ~ bs + lenroll + lstaff+ lunch, data = benefits)
summary(modelo)
```

```
##
## Call:
## lm(formula = lavgsal ~ bs + lenroll + lstaff + lunch, data = benefits)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26047 -0.10976 -0.00849  0.10368  0.59383
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.7236149  0.1121095 122.413  < 2e-16 ***
## bs          -0.1774396  0.1219691  -1.455 0.145897
## lenroll     -0.0292406  0.0084997  -3.440 0.000594 ***
## lstaff      -0.6907025  0.0184598 -37.417  < 2e-16 ***
## lunch       -0.0008471  0.0001625  -5.213 2.07e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.1677 on 1843 degrees of freedom
## Multiple R-squared:  0.4826, Adjusted R-squared:  0.4815
## F-statistic: 429.8 on 4 and 1843 DF,  p-value: < 2.2e-16
```

- Quais variáveis são estatísticamente significativas? lenroll, lstaff, lunch
- Interprete os betas

| Modelo | V. Dep | V. Indep | Interpretação $\beta_1$ |
|---|---|---|---|
| Nível-Nível | $y$ | $x$ | $\Delta y = \beta_1 \Delta x$ |
| Nível-Log | $y$ | $\log(x)$ | $\Delta y = (\beta_1/100)\%\Delta x$ |
| Log-Nível | $\log(y)$ | $x$ | $\%\Delta y = 100\beta_1 \Delta x$ |
| Log-Log | $\log(y)$ | $\log(x)$ | $\%\Delta y = \beta_1 \%\Delta x$ |

```
summary(modelo)$coef
```

```
##                  Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept) 13.7236149410 0.1121095023 122.412594  0.000000e+00
## bs          -0.1774396290 0.1219690987  -1.454792  1.458972e-01
## lenroll     -0.0292405955 0.0084997318  -3.440179  5.942851e-04
## lstaff      -0.6907024921 0.0184598195 -37.416535 1.974953e-228
## lunch       -0.0008470929 0.0001624916  -5.213150  2.065104e-07
```

- Com o aumento de enroll em 1% espera-se uma diminuição em avgsal de 0.029%
- Com o aumento de staff em 1% espera-se uma diminuição em avgsal de 0.691%
- Com o aumento de uma unidade em lunch, espera-se que avgsal diminue em 0.08%

- Calcule intervalos de confiança 95% para os betas

```
confint(modelo, level = 0.95)
```

```
##                   2.5 %       97.5 %
## (Intercept) 13.50373996 13.943489926
## bs          -0.41665177  0.061772509
## lenroll     -0.04591071 -0.012570480
## lstaff      -0.72690685 -0.654498134
## lunch       -0.00116578 -0.000528406
```

- adicione a variável $lunch^2$ e verifique se a qualidade do ajuste do medelo melhorou ou não.

```
modelo2 = lm(lavgsal ~ bs + lenroll + lstaff+ lunch + I(lunch^2), data = benefits)
summary(modelo1)$adj.r.squared
```

```
## [1] 0.3619087
```

```
summary(modelo2)$adj.r.squared # melhorou
```

```
## [1] 0.489761
```

Para exercícios adicionais, veja as Seções **Exercícios em computador** dos Capítulos 4, 6 e 7 do livro texto.