More on Linear Regression

Stat 120

May 29 2023

Simple Linear Regression: Linear Regression of BAC (y) on Beers (x)

$$\hat{\mu}(BAC \mid X) = -0.0127 + 0.0180(\text{ Beers })$$
$$\hat{\sigma} = 0.02044$$

```
Call:
lm(formula = BAC ~ Beers, data = bac)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.027118 -0.017350  0.001773  0.008623  0.041027 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.012701   0.012638  -1.005    0.332    
Beers        0.017964   0.002402   7.480 2.97e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02044 on 14 degrees of freedom
Multiple R-squared:  0.7998,	Adjusted R-squared:  0.7855 
F-statistic: 55.94 on 1 and 14 DF,  p-value: 2.969e-06
```

Inference using R-output

Inference uses t-distributions with df = n-2

Test: Does X have an effect on the mean of Y?

Hypotheses:

$$\begin{aligned} \mathrm{H}_0 :& \beta_i = 0 && \text{(no effect for predictor i)} \\ \mathrm{H}_A :& \beta_i \neq 0 && \text{(predictor i has an effect on y)} \end{aligned}$$

Test Stat: labeled t-value in R output

$$t = \frac{\hat{\beta}_i - \mathbf{0}}{\mathrm{SE}\left(\hat{\beta}_i\right)}$$

P-value: two-tailed, label "Pr $(> |\mathrm{t}|)$ " in R output

Confidence Interval

C% confidence interval for β_i *is* $\hat{\beta}_i \pm t^* \, \mathrm{SE}\left(\hat{\beta}_i\right)$

- *Get CIs for slope/intercept with* `confint` *command or compute using* `qt(.975, df= )` *to get t* for 95% CI*

```
confint(bac.lm)  # 95% C.I.
                  2.5 %     97.5 %
(Intercept) -0.03980535 0.01440414
Beers        0.01281262 0.02311490
```

Inference for slope (effect of Beers on BAC)

$$
\begin{aligned}
\mathrm{H}_0 : \beta_i &= 0 && \text{(no effect for predictor i)} \\
\mathrm{H}_A : \beta_i &\neq 0 && \text{(predictor i has an effect on y)}
\end{aligned}
$$

term	estimate	std.error	statistic	p.value
(Intercept)	-0.0127006	0.0126375	-1.004993	0.3319551
Beers	0.0179638	0.0024017	7.479592	0.0000030

(a) For this example, the slope test statistic is

$$
\mathrm{t} = \frac{(0.018 - 0)}{0.0024} = 7.48
$$

and the p-value is less than 0.0001**.**

(b) Inference for slope (effect of Beers on BAC)

- *The observed slope of* 0.018 *is* 7.48SE *'s away from the hypothesized slope of 0 .*
- *If we repeated this experiment many times, less than* 0.01% *of the time we would see an observed slope that is* 7.48SE *'s or more away from 0 if the true slope was 0.*
- *The effect of the number of beers on* BAC *is statistically significant* $(t = 7.48, \text{df} = 14, p < 0.0001)$.

How much of an effect?

(c) Inference for slope (effect of Beers on BAC)

95% CI for true slope (effect on the mean):

```r
bac <- read.csv("https://raw.githubusercontent.com/deepbas/statdatasets/main/BAC.csv")
bac.lm <- lm(BAC ~ Beers, data = bac) # fit the model
knitr::kable(confint(bac.lm)) # confidence interval
```

	2.5 %	97.5 %
(Intercept)	-0.0398054	0.0144041
Beers	0.0128126	0.0231149

$$0.018 \pm 2.1448(0.0024) = (0.013, 0.023)$$

Each additional beer is associated with an average increase in BAC of 0.018 (95% CI 0.013 to 0.023).

Inference for intercept (predicted BAC at 0 beers)

$$\begin{aligned} \mathrm{H}_0 : &\beta_0 = 0 & \text{(true (population) intercept} = 0) \\ \mathrm{H}_A : &\beta_0 \neq 0 & \text{(true (population) intercept is not 0)} \end{aligned}$$

term	estimate	std.error	statistic	p.value
(Intercept)	-0.0127006	0.0126375	-1.004993	0.3319551
Beers	0.0179638	0.0024017	7.479592	0.0000030

For this example, the intercept test statistic is

$$\mathrm{t} = \frac{(-0.0127 - 0)}{0.0126} = -1.005$$

and the p**-value is** 0.332**. The predicted BAC at 0 beers is not significantly different from 0** $(\mathrm{t} = -1.01, \mathrm{df} = 14, \mathrm{p}$**-value** $= 0.33)$**.**

Summary

- Number of beers has a statistically significant effect on BAC.
- We can explain about 80% of the variation in BAC by knowing the number of beers drank.
- How can we explain the other 20% of variation in BAC observed in this data?
- Multiple regression: include more predictors (explanatory variables) of BAC.

Multiple Linear Regression Model

Suppose you have p explanatory variables

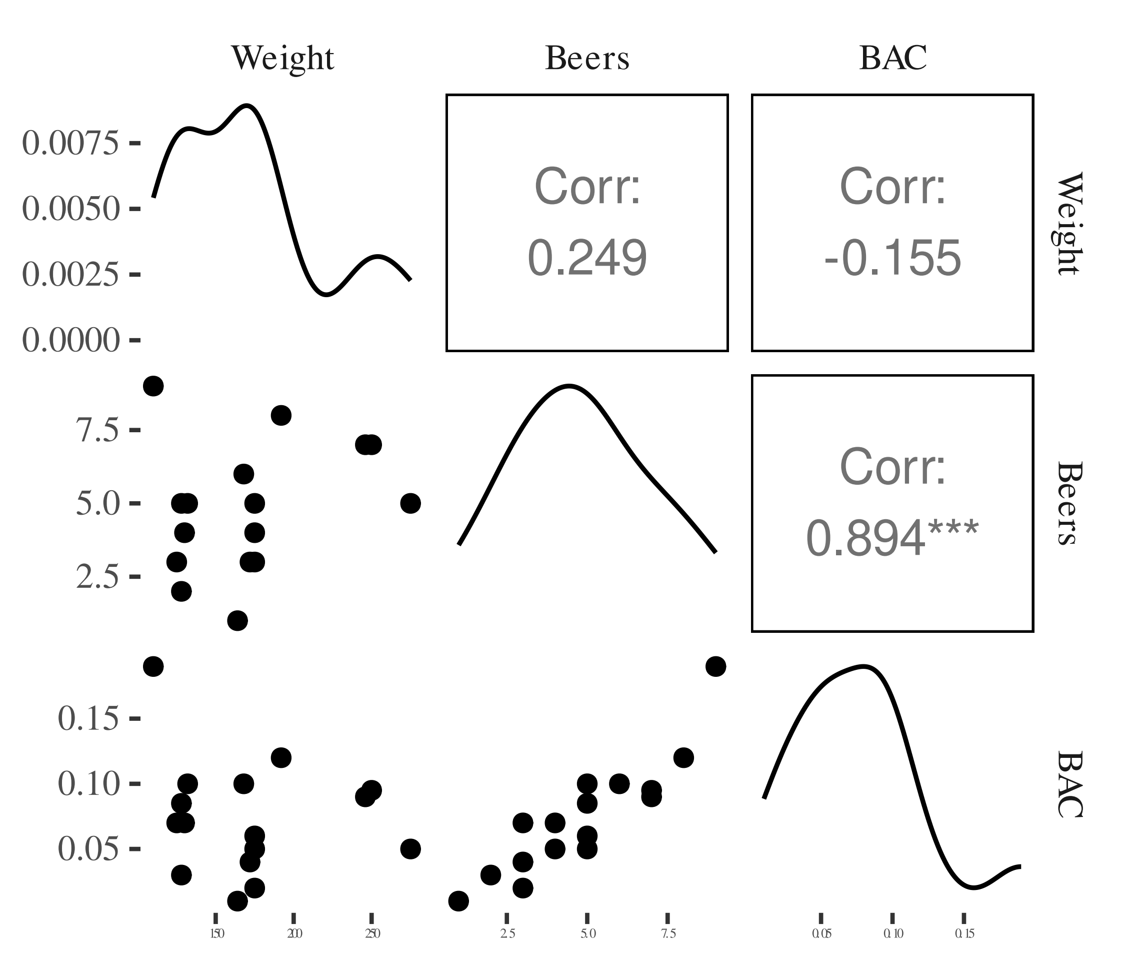
- ***Mean of Y is a linear function of $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$***

  $$\mu(Y \mid X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

- ***Model SD is $\boldsymbol{\sigma}$ for all values of X.***

Scatterplot matrix for BAC example

```
library(GGally)
ggpairs(bac[, -c(1,2,3)]) + theme(axis.text.x = element_text(size = 3))
```



Regression of BAC on Beers and Weight

```
Call:
lm(formula = BAC ~ Beers + Weight, data = bac)

Residuals:
       Min         1Q     Median         3Q        Max 
-0.0162968 -0.0067796  0.0003985  0.0085287  0.0155621 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.986e-02  1.043e-02   3.821  0.00212 ** 
Beers        1.998e-02  1.263e-03  15.817 7.16e-10 ***
Weight      -3.628e-04  5.668e-05  -6.401 2.34e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01041 on 13 degrees of freedom
Multiple R-squared:  0.9518,    Adjusted R-squared:  0.9444 
F-statistic: 128.3 on 2 and 13 DF,  p-value: 2.756e-09
```

(d) The fitted model for BAC is

$$\widehat{BAC} = \hat{\mu}(BAC \mid X) = 0.0399 + 0.0200(Beers) - 0.00036(Weight).$$

Knowing number of beers and weight allows us to explain about 95% of the observed variation in BAC levels

Interpreting parameters in $\mu(Y \mid X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$

The fitted model for BAC is

$$\widehat{BAC} = \hat{\mu}(BAC \mid X) = 0.0399 + 0.0200(Beers) - 0.00036(Weight).$$

(e) John: weighs 160 lbs, drank 4 beers

$$\widehat{BAC} = 0.0399 + 0.0200(4) - 0.00036(160) = 0.0623$$

(e) Bob: weighs 160 lbs, drank 5 beers

$$\widehat{BAC} = 0.0399 + 0.0200(5) - 0.00036(160) = 0.0823$$

(e) Difference between John and Bob's predicted BAC?

$$0.0823 - 0.0623 = 0.0200$$

Inference

$$H_0 : \beta_i = 0 \qquad \text{(no effect for predictor i)} \qquad t = \frac{\hat{\beta}_i - 0}{\mathrm{SE}\left(\hat{\beta}_i\right)}$$
$$H_A : \beta_i \neq 0$$

- ***Inference (p-values, CI) based on t-distribution with*** $\mathrm{df} = \mathrm{n} - (\mathrm{p} + 1) = \mathrm{n} - (\text{betas in model})$
- ***R-squared: proportion of variation in y explained by the model***
- ***Test: Is X_i needed in a model that already contains all other predictors?***
- ***C% confidence interval for β_i is*** $\hat{\beta}_i \pm t^{*}\mathbf{SE}\left(\hat{\beta}_i\right)$

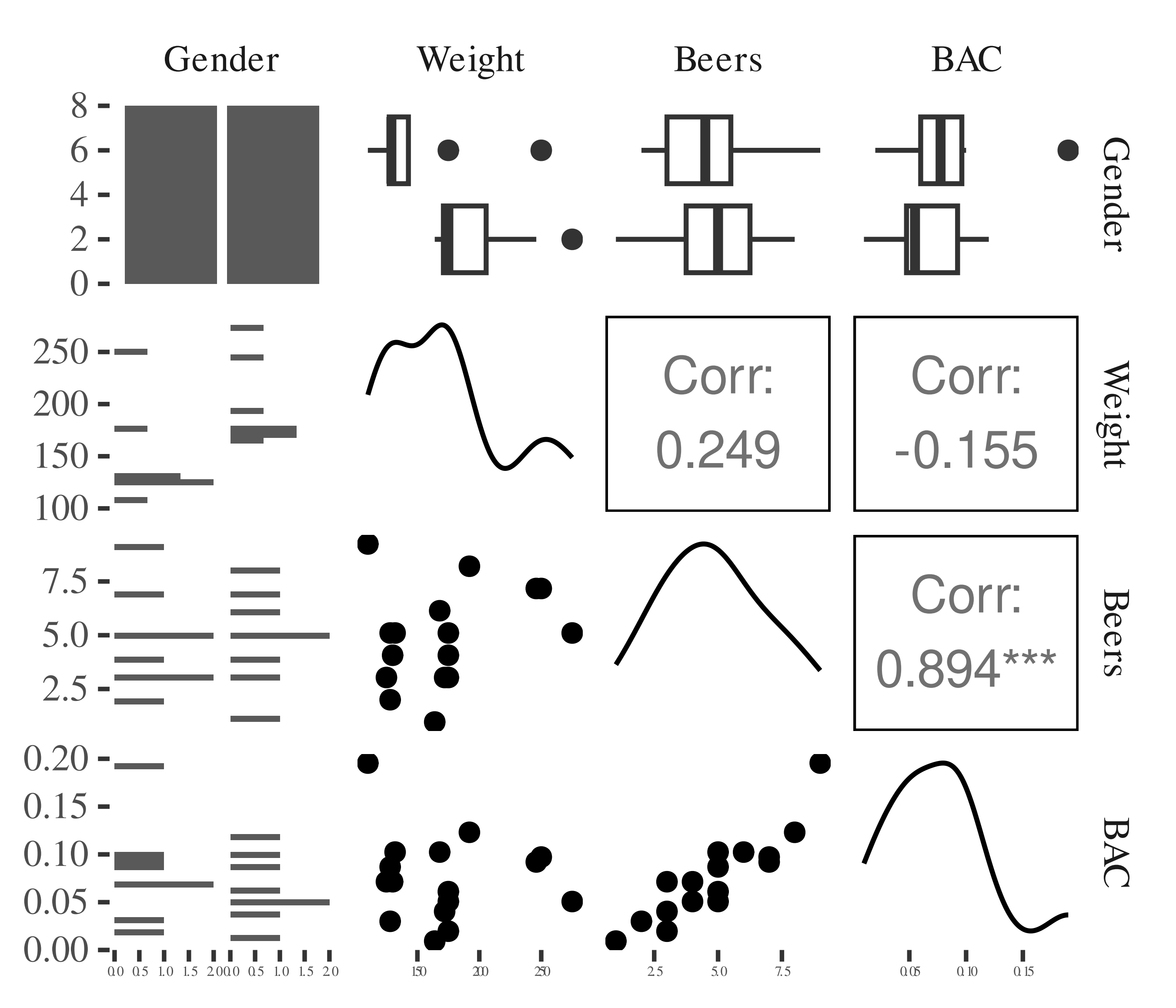
Regression of BAC on Beers and Weight

term	estimate	std.error	statistic	p.value
(Intercept)	0.0398634	0.0104333	3.820787	0.0021219
Beers	0.0199757	0.0012629	15.817343	0.0000000
Weight	-0.0003628	0.0000567	-6.401230	0.0000234

	2.5 %	97.5 %
(Intercept)	0.0173236	0.0624031
Beers	0.0172474	0.0227040
Weight	-0.0004853	-0.0002404

(h) Both number of beers and weight are statistically significant predictors of BAC (p-value < 0.0001). Holding weight constant, we are 95% confident that the true effect of drinking one more beer is a 0.017 to 0.23 unit increase in mean BAC.

Scatterplot Matrix for BAC example



Regression of BAC on Beers, Weight and Gender

```r
lm2.bac <- lm(BAC ~ Beers + Weight + Gender, data = bac) # fit the model
summary(lm2.bac)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.871e-02  1.097e-02   3.528 0.004164 **
Beers        1.990e-02  1.309e-03  15.196 3.35e-09 ***
Weight      -3.444e-04  6.842e-05  -5.034 0.000292 ***
Gendermale  -3.240e-03  6.286e-03  -0.515 0.615584
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01072 on 12 degrees of freedom
Multiple R-squared:  0.9528,    Adjusted R-squared:  0.941
F-statistic: 80.81 on 3 and 12 DF,  p-value: 3.162e-08
```

The fitted model for BAC is

$$\widehat{BAC} = 0.039 + 0.020(Beers) - 0.00034(Weight) - 0.0032\ (\text{Male})$$

Regression of BAC on Beers, Weight and Gender

$$\widehat{BAC} = 0.039 + 0.020(Beers) - 0.00034(Weight) - 0.0032\ (\text{Male})$$

"Male" is an indicator variable that equals 1 when we want to predict male BAC **and 0 when we want to predict Female BAC.**

(i) Barb drank 4 beers, weighs 160 lbs and is female:

$$\widehat{BAC} = 0.039 + 0.020(4) - 0.00034(160) - 0.0032(0) = 0.0646$$

(i) John drank 4 beers, weighs 160 lbs and is male:

$$\widehat{BAC} = 0.039 + 0.020(4) - 0.00034(160) - 0.0032(1) = 0.0614$$

Regression of BAC on Beers, Weight and Gender

How is the effect of GenderMale -0.0032 interpreted?

(j) Holding weight and beers constant, we predict that the BAC of males is 0.0032 units lower than females.

Regression of BAC on Beers, Weight and Gender

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.871e-02  1.097e-02   3.528 0.004164 **
Beers        1.990e-02  1.309e-03  15.196 3.35e-09 ***
Weight      -3.444e-04  6.842e-05  -5.034 0.000292 ***
Gendermale  -3.240e-03  6.286e-03  -0.515 0.615584
```

(k) But, is the effect of Gender statistically significant?

- No - the p-value of 0.6155 shows that there is no statistically significant difference between the mean BAC of males and females after accounting for their weight and number of beers drank!

There is a lot more to cover ..

Model diagnostic tools

- *Are model conditions met?*
- *linearity, constant variance, independent observation*

Outlier checks

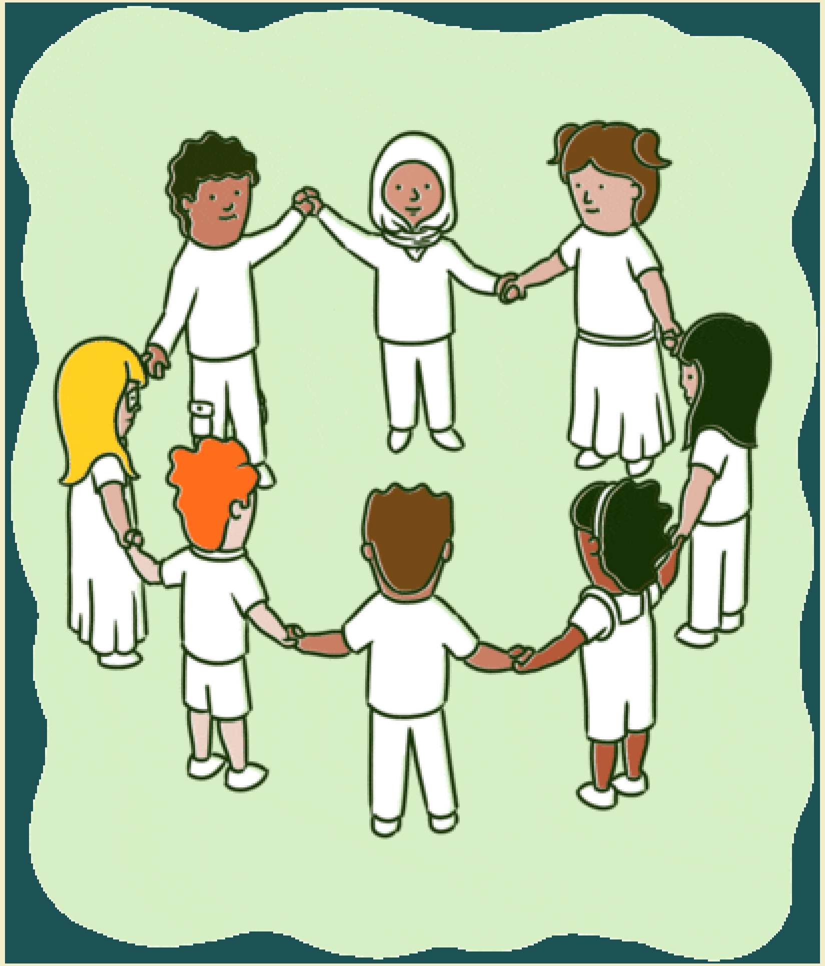
- *What is an "outlier" in multiple regression?*
- *Can we determine how influential a case is?*

More sophisticated models

- *Transformations (logs, square roots)*
- *Predictor interactions (does the effect of weight on BAC depend on gender?)*
- *Logistic regression: response is categorical!*

YOUR TURN 1

05:00



- *Go over to the in class activity file*
- *Complete the remaining activity*