Linear regression

Advanced statistical methods and models in experimental design

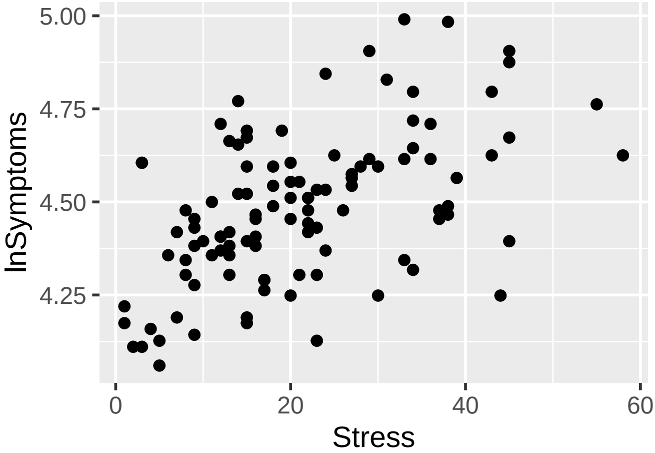
Bartosz Maćkiewicz

Stress and mental health (`wagner1988.csv`)

Wagner, Compas, and Howell (1988) explored the relationship between stress and mental health among first-year college students. They developed a scale to assess the frequency, perceived significance, and desirability of recent life events. This scale helped to create a measure of negative events, weighted by their reported frequency and the respondents' subjective assessment of each event's impact. This metric served to quantify the subjects' perceived social and environmental stress. Additionally, the students completed the Hopkins Symptom Checklist, which evaluates the presence or absence of 57 psychological symptoms.

Stress and mental health (wagner1988.csv)

```
wagner <- read.csv("wagner1988.csv")
library(ggplot2)
ggplot(data = wagner) +
  aes(x = Stress, y = lnSymptoms) +
  geom_point()
```



The Regression Line

Recall from high school mathematics that the equation of a straight line is expressed as $Y = bX + a$. For analytical purposes, we represent this equation as:

$$\hat{Y} = bX + a$$

Where:

- $\hat{Y}$ = the predicted value of Y
- b = the slope of the regression line (the amount of difference in $\hat{Y}$ associated with a one-unit difference in X)
- a = the intercept (the value of $\hat{Y}$ when $X = 0$)
- X = the value of the predictor variable

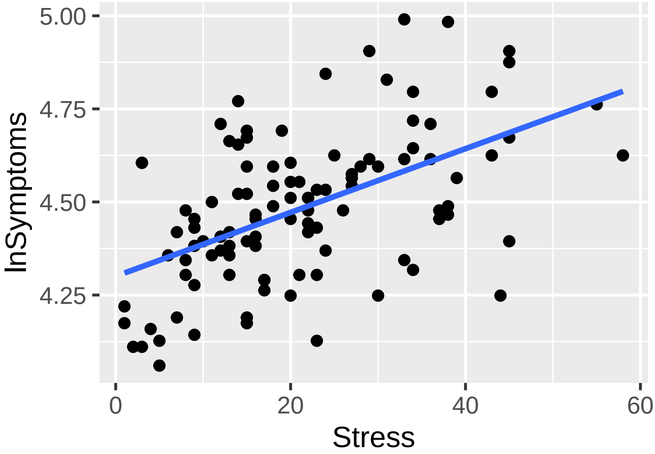
The Regression Line: best fitting line

Our objective is to determine the values of a and b that yield the **best-fitting** linear function. This means we aim to align the regression line —- represented by the values of $\hat{Y}$ for various X values —- as closely as possible with the actual observed values of Y.

The critical question then arises: How do we define “best-fitting”?

The Regression Line: best fitting line

```r
ggplot(data = wagner) +
  aes(x = Stress, y = lnSymptoms) +
  geom_point() +
  geom_smooth(method = lm, se = F) # add regression line
```



The Regression Line: errors of prediction

A logical way would be in terms of **errors of prediction** — that is, in terms of the $(Y - \hat{Y})$ deviations.

Because $\hat{Y}$ is the value of the symptom (`lnSymptoms`) variable that our equation would predict for a given level of stress, and Y is a value that we actually obtained, $(Y - \hat{Y})$ is the error of prediction, usually called the residual.

We want to find the line (the set of $\hat{Y}$s) that minimizes such errors. We cannot just minimize the sum of the errors, however, because for an infinite variety of lines—any line that goes through the point (X, Y) - that sum will always be zero. (We will overshoot some and undershoot others.)

Instead, we will look for that line that minimizes the sum of the squared errors - that minimizes $\sum(Y - \hat{Y})^2$.

The Regression Line: errors of prediction

```r
library(tigerstats)
#shiny::runApp(system.file("FindRegLine", package="tigerstats"))
```

The Regression Line: `lm` function

```r
fit <- lm(lnSymptoms ~ Stress, data = wagner)
summary(fit)
```

```
##
## Call:
## lm(formula = lnSymptoms ~ Stress, data = wagner)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42889 -0.13568  0.00478  0.09672  0.40726
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.300537   0.033088 129.974  < 2e-16 ***
## Stress      0.008565   0.001342   6.382 4.83e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1726 on 105 degrees of freedom
## Multiple R-squared:  0.2795, Adjusted R-squared:  0.2726
## F-statistic: 40.73 on 1 and 105 DF,  p-value: 4.827e-09
```

Interpretations of Regression. Intercept

- We have defined the intercept as that value of $\hat{Y}$ when X equals zero. As such, it has meaning in some situations and not in others, primarily depending on whether or not $X = 0$ has meaning and is near or within the range of values of X used to derive the estimate of the intercept.
- Centering the data. In many situations, it is useful to "center" your data at the mean by subtracting the mean of X from every X value. If you do this, an X value of 0 now represents the mean X and the intercept is now the value predicted for Y when X is at its mean.

Interpretations of Regression. Intercept. Centering the data

```r
wagner$StressC <- wagner$Stress - mean(wagner$Stress)
fit <- lm(lnSymptoms ~ StressC, data = wagner)
summary(fit)
```

```
##
## Call:
## lm(formula = lnSymptoms ~ StressC, data = wagner)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42889 -0.13568  0.00478  0.09672  0.40726
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.482879   0.016686 268.663  < 2e-16 ***
## StressC     0.008565   0.001342   6.382 4.83e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1726 on 105 degrees of freedom
## Multiple R-squared:  0.2795, Adjusted R-squared:  0.2726
## F-statistic: 40.73 on 1 and 105 DF,  p-value: 4.827e-09
```

Interpretations of Regression. Slope

We have defined the slope (regression coefficient) as the change in $\hat{Y}$ for a one-unit change in X. As such it is a measure of the predicted rate of change in Y. By definition, then, the slope is often a meaningful measure.

Standardized coefficients

Although we rarely work with standardized data (data that have been transformed so as to have a mean of zero and a standard deviation of one on each variable), it is worth considering what b would represent if the data for each variable were standardized separately.

In that case, a difference of one unit in X or Y would represent a difference of one standard deviation. Thus, if the slope was 0.75, for standardized data, we could say that a one standard deviation increase in X will be reflected in three-quarters of a standard deviation increase in $\hat{Y}$. When speaking of the slope for standardized data, we often refer to the standardized regression coefficient as β (*beta*) to differentiate it from the coefficient for nonstandardized data (b).

Standardized coefficients

```r
wagner$StressS <- wagner$StressC /  sd(wagner$StressC)
wagner$lnSymptomsS <- (wagner$lnSymptoms - mean(wagner$lnSymptoms)) /
sd(wagner$lnSymptoms)
fit <- lm(lnSymptomsS ~ StressS, data = wagner)
summary(fit)
```

```
## 
## Call:
## lm(formula = lnSymptomsS ~ StressS, data = wagner)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2.11929 -0.67043  0.02364  0.47792  2.01238 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -7.071e-16  8.245e-02   0.000        1    
## StressS      5.287e-01  8.284e-02   6.382 4.83e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.8529 on 105 degrees of freedom
## Multiple R-squared:  0.2795, Adjusted R-squared:  0.2726
```

Standardized coefficients

```r
library(lm.beta)
fit <- lm(lnSymptoms ~ Stress, data = wagner)
summary(lm.beta(fit))
```

```
## 
## Call:
## lm(formula = lnSymptoms ~ Stress, data = wagner)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.42889 -0.13568  0.00478  0.09672  0.40726 
## 
## Coefficients:
##             Estimate Standardized Std. Error t value Pr(>|t|)    
## (Intercept) 4.300537           NA   0.033088 129.974  < 2e-16 ***
## Stress      0.008565     0.528656   0.001342   6.382 4.83e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.1726 on 105 degrees of freedom
## Multiple R-squared:  0.2795, Adjusted R-squared:  0.2726 
## F-statistic: 40.73 on 1 and 105 DF,  p-value: 4.827e-09
```

Hypothesis testing in linear regression

It can be shown that b is normally distributed about $b*$ (population b) with a standard error approximated by:

$$s_b = \frac{s_{Y \cdot X}}{s_X \sqrt{N-1}}$$

Thus, if we wish to test the hypothesis that the true slope of the regression line in the population is zero (H_0: $b* = 0$) , we can simply form the ratio:

$$t = \frac{b - b*}{s_b} = \frac{b}{\frac{s_{Y \cdot X}}{s_X \sqrt{N-1}}} = \frac{(b)(s_X)(\sqrt{N-1})}{s_{Y \cdot X}}$$

which is distributed as t on $N - 2$ df

Hypothesis testing in linear regression

```
predicted <- wagner$Stress * 0.008565 + 4.300537
ss_res <- sum((wagner$lnSymptoms - predicted)**2)
s_X_dot_Y <- sqrt(ss_res / (nrow(wagner) - 2))
s_b <- s_X_dot_Y / (sd(wagner$Stress) * sqrt(nrow(wagner) - 1))
t <- (0.008565 - 0) / s_b
t
## [1] 6.381966
pt(t, nrow(wagner) - 2, lower.tail = F) * 2
## [1] 4.824136e-09
```