How well does the model fit the data?

Goodness of fit metrics

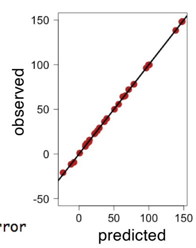
1

$$Y = 42\text{cm} + 2.3X_1 + 11X_2 + e$$

The variance of residuals in relation to the variance of predictors regulates how well the model fits the data

$e \sim N(0, \sigma^2) \therefore e \sim N(0,1)$

```
n = 30
constant = 42
X1 = rnorm(n,1,4)
X2 = rnorm(n,1,4)
error = rnorm(n,0,1)

Y = constant + 2.3*X1 + 11*X2 + error
```

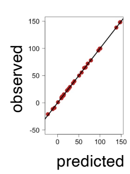


2

$$Y = 42\text{cm} + 2.3X_1 + 11X_2 + e$$

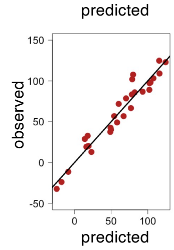
```
n = 30
constant = 42
X1 = rnorm(n,1,4)
X2 = rnorm(n,1,4)
error = rnorm(n,0,1)
```

$e \sim N(0, \sigma^2)$
$e \sim N(0,1)$




```
n = 30
constant = 42
X1 = rnorm(n,1,4)
X2 = rnorm(n,1,4)
error = rnorm(n,0,10)

Y = constant + 2.3*X1 + 11*X2 + error
```

$e \sim N(0, \sigma^2)$
$e \sim N(0,100)$

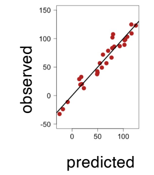


3

$Y = 42\text{cm} + 2.3X_1 + 11X_2 + e$

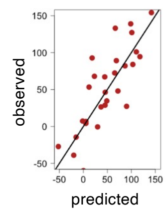
```
n = 30
constant = 42
X1 = rnorm(n,1,4)
X2 = rnorm(n,1,4)
error = rnorm(n,0,10)
```

$e \sim N(0, \sigma^2)$
$e \sim N(0{,}100)$




```
n = 30
constant = 42
X1 = rnorm(n,1,4)
X2 = rnorm(n,1,4)
error = rnorm(n,0,40)
```

$e \sim N(0, \sigma^2)$
$e \sim N(0{,}1600)$

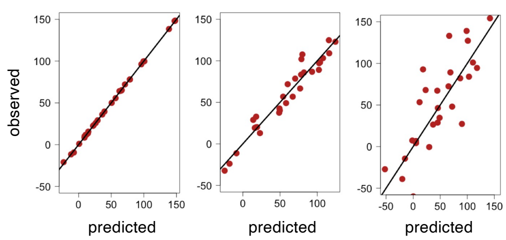



```
Y = constant + 2.3*X1 + 11*X2 + error
```

4

$Y = 42\text{cm} + 2.3X_1 + 11X_2 + e$

Same model, increase in error (residual variation)



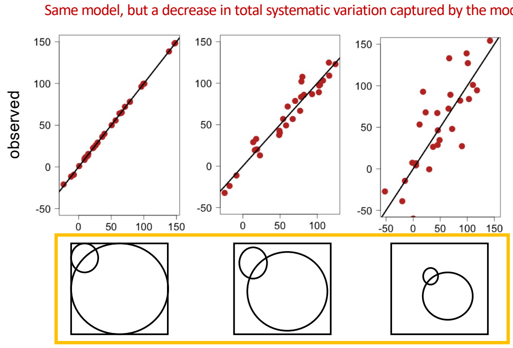
$e \sim N(0{,}1)$ $e \sim N(0{,}100)$ $e \sim N(0{,}1600)$

5

$Y = 42\text{cm} + 2.3X_1 + 11X_2 + e$

Same model, increase in error (residual variation)

Same model, but a decrease in total systematic variation captured by the model

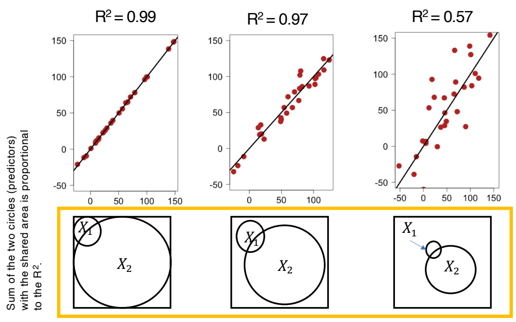


6

Assessing how well the model fit the data – Goodness of fit metrics

1) Coefficient of determination (R^2) – a measure of how well the estimated regression line approximates the observed data points. It is often interpreted as the percentage of total variation explained by the regression model.

$$Y = 42\text{cm} + 2.3X_1 + 11X_2 + e$$



7

Assessing how well the model fit the data – Goodness of fit metrics

1) Coefficient of determination (R^2) – a measure of how well the estimated regression line approximates the observed data points. It is often interpreted as the percentage of total variation explained by the regression model.

It can be calculated in many ways (always leading to the same result), but here are three of them (no need to memorize them):

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \hat{Y})^2}{\sum_{i=1}^{n}(Y_i - Y)^2} = \frac{total\ SS\ predicted}{total\ SS\ observed}$$

8

Assessing how well the model fit the data – Goodness of fit metrics

1) Coefficient of determination (R^2) – a measure of how well the estimated regression line approximates the observed data points. It is often interpreted as the percentage of total variation explained by the regression model.

It can be calculated in many ways (always leading to the same result), but here are three of them:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \hat{Y})^2}{\sum_{i=1}^{n}(Y_i - Y)^2} = \frac{total\ SS\ predicted}{total\ SS\ observed}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(Y_i - Y)^2} = 1 - \frac{total\ SS\ error\ (residuals)}{total\ SS\ observed}$$

9

Assessing how well the model fit the data – Goodness of fit metrics

1) Coefficient of determination (R^2) – a measure of how well the estimated regression line approximates the observed data points. It is often interpreted as the percentage of total variation explained by the regression model.

It can be calculated in many ways (always leading to the same result), but here are three of them:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{total\ SS\ predicted}{total\ SS\ observed}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{total\ SS\ error\ (residuals)}{total\ SS\ observed}$$

$$R^2 = cor(observed, predicted)^2$$

10

Assessing how well the model fit the data – Goodness of fit metrics

```
> lm.res = lm(Y~X1+X2)
> summary(lm.res)

Call:
lm(formula = Y ~ X1 + X2)

Residuals:
    Min      1Q  Median      3Q     Max
-69.871 -25.949  -2.132  23.879 103.969

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   44.243      8.966   4.935 3.63e-05 ***
X1             3.561      1.793   1.986   0.0572 .
X2            10.177      1.782   5.711 4.54e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 42.89 on 27 degrees of freedom
Multiple R-squared:  0.565,	Adjusted R-squared:  0.5328
F-statistic: 17.54 on 2 and 27 DF,  p-value: 1.316e-05
```

11

What happens to the R^2 when non-relevant predictors are considered in the model?

True population model
(i.e., only two relevant predictors)

$$Y = 42\text{cm} + 2.3X_1 + 11X_2 + e$$

Previous model just with the two relevant predictors:

```
Residual standard error: 42.89 on 27 degrees of freedom
Multiple R-squared:  0.565,	Adjusted R-squared:  0.5328
F-statistic: 17.54 on 2 and 27 DF,  p-value: 1.316e-05
```

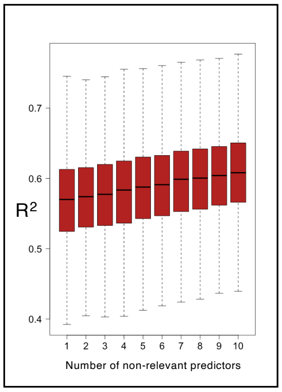
Previous model with the two relevant predictors plus two irrelevant predictors X3 and X4:

```
n = 30
X3=rnorm(n,1,4)
X4=rnorm(n,1,4)
lm.res = lm(Y~X1+X2+X3+X4)
summary(lm.res)
```

```
> lm.res = lm(Y~X1+X2+X3+X4)
> summary(lm.res)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4)

Residuals:
    Min      1Q  Median      3Q     Max
-59.933 -28.631  -2.034  28.055  85.368

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   40.846      9.006   4.535 0.000124 ***
X1             3.600      1.755   2.051 0.050874 .
X2             9.592      1.759   5.452 1.16e-05 ***
X3             1.349      2.077   0.649 0.522086
X4             3.164      2.047   1.545 0.134861
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 41.59 on 25 degrees of freedom
Multiple R-squared:  0.6212,	Adjusted R-squared:  0.5606
F-statistic: 10.25 on 4 and 25 DF,  p-value: 4.708e-05
```

12

What happens to the R^2 when non-relevant predictors are considered in the model?



Simulation with 1000 samples (n=100) from model:

$$Y = 42\text{cm} + 2.3X_1 + 11X_2 + e$$

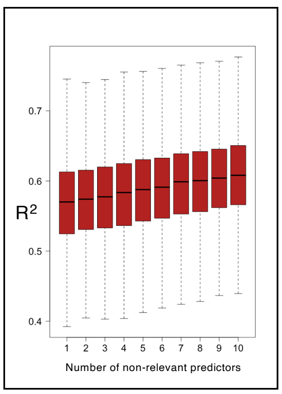
adding from 1 to 10 non-relevant predictors

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{total\ SS\ predicted}{total\ SS\ observed}$$

As the number of predictors increase, it is more likely that they will improve the model even by pure chance (i.e., non-relevant predictors)

13

What happens to the R^2 when non-relevant predictors are considered in the model?



Simulation with 1000 samples (n=100) from model:

$$Y = 42\text{cm} + 2.3X_1 + 11X_2 + e$$

adding from 1 to 10 non-relevant predictors

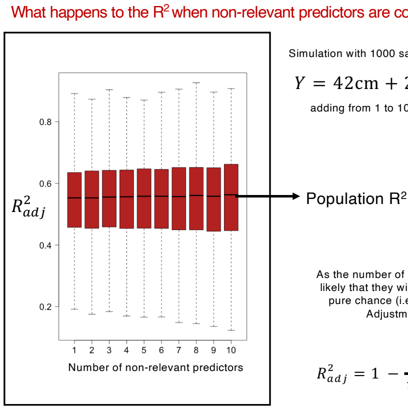
$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{total\ SS\ predicted}{total\ SS\ observed}$$

As the number of predictors increase, it is more likely that they will improve the model even by pure chance (i.e., non-relevant predictors)

```
Residual standard error: 42.89 on 27 degrees of freedom
Multiple R-squared:  0.565,     Adjusted R-squared:  0.5328
F-statistic: 17.54 on 2 and 27 DF,  p-value: 1.316e-05
```

14

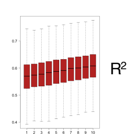
What happens to the R^2 when non-relevant predictors are considered in the model?



Simulation with 1000 samples (n=100) from model:

$$Y = 42\text{cm} + 2.3X_1 + 11X_2 + e$$

adding from 1 to 10 non-relevant predictors



As the number of predictors increase, it is more likely that they will improve the model even by pure chance (i.e., non-relevant predictors). Adjustments are necessary:

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}(1 - R^2)$$

note that accuracy is great but precision is reduced as the number of predictors increases

15

A complete empirical example

16

Empirical example:
Understanding the drivers of pollution in US cities

17

Empirical example - What are the drivers of air pollution (sulfur dioxide) in US cities?

A	B	C	D	E	F	G	H
city	SO2	temp	manu	popul	wind	precip	raindays
Albany	46	47.6	44	116	8.8	33.36	135
Albuquerque	11	56.8	46	244	8.9	7.77	58
Atlanta	24	61.5	368	497	9.1	48.34	115
Baltimore	47	55	625	905	9.6	41.31	111
Buffalo	11	47.1	391	463	12.4	36.11	166
Charleston	31	55.2	35	71	6.5	40.75	148
Chicago	110	50.6	3344	3369	10.4	34.44	122
Cincinnati	23	54	462	453	7.1	39.04	132
Cleveland	65	49.7	1007	751	10.9	34.99	155
Columbus	26	51.5	266	540	8.6	37.01	134
Dallas	9	66.2	641	844	10.9	35.94	78
Denver	17	51.9	454	515	9	12.95	86
Des Moines	17	49	104	201	11.2	30.85	103
Detroit	35	49.9	1064	1513	10.1	30.96	129
Hartford	56	49.1	412	158	9	43.37	127
Houston	10	68.9	721	1233	10.8	48.19	103
Indianapolis	28	52.3	361	746	9.7	38.74	121
Jacksonville	14	68.4	136	529	8.8	54.47	116
Kansas City	14	54.5	381	507	10	37	99
Little Rock	13	61	91	132	8.2	48.52	100
Louisville	30	55.6	291	593	8.3	43.11	123
Memphis	10	61.6	337	624	9.2	49.1	105
Miami	10	75.5	207	335	9	59.8	128
Milwaukee	16	45.7	569	717	11.8	29.07	123
Minneapolis	29	43.5	699	744	10.6	25.94	137
Nashville	18	59.4	275	448	7.9	46	119
New Orleans	9	68.3	204	361	8.4	56.77	113
Norfolk	31	59.3	96	308	10.6	44.68	116
Omaha	14	51.5	181	347	10.9	30.18	98
Philadelphia	69	54.6	1692	1950	9.6	39.93	115
Phoenix	10	70.3	213	582	6	7.05	36
Pittsburgh	61	50.4	347	520	9.4	36.22	147
Providence	94	50	343	179	10.6	42.75	125
Richmond	26	57.8	197	299	7.6	42.59	115
Salt Lake City	28	51	137	176	8.7	15.17	89
San Francisco	12	56.7	453	716	8.7	20.66	67
Seattle	29	51.1	379	531	9.4	38.79	164

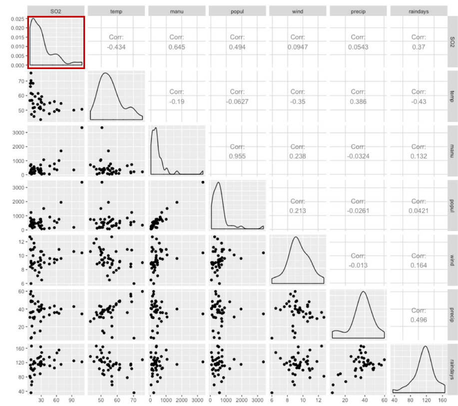
⋮

n=41

- City: City
- SO_2: Sulfur dioxide content of air in micrograms per cubic meter.
- Temp: Average annual temperature in degrees Fahrenheit.
- Manu: Number of manufacturing enterprises employing 20 or more workers.
- Popul: Population size in thousands from the 1970 census.
- Wind: Average annual wind speed in miles per hour.
- Precip: Average annual precipitation in inches.
- Raindays: Average number of days with precipitation per year.

18

First step: Look at the raw data (distribution and pairwise correlations)



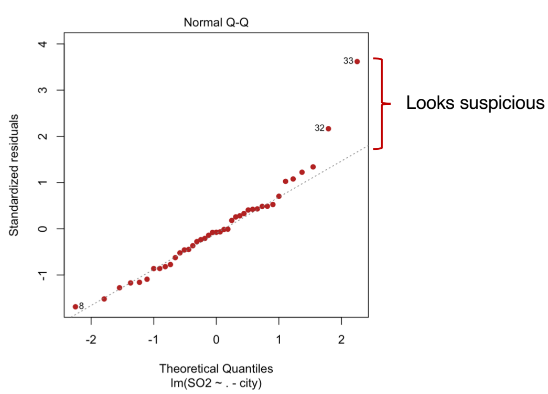
19

Second step: Run the model

```
> fit <- lm(SO2 ~ temp + manu + popul + wind + precip + raindays , data=data.pollution)
> summary(fit)

Call:
lm(formula = SO2 ~ temp + manu + popul + wind + precip + raindays,
    data = data.pollution)

Residuals:
    Min      1Q  Median      3Q     Max
-23.004  -8.542  -0.991   5.758  48.758

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 111.72848   47.31810   2.361 0.024087 *
temp         -1.26794    0.62118  -2.041 0.049056 *
manu          0.06492    0.01575   4.122 0.000228 ***
popul        -0.03928    0.01513  -2.595 0.013846 *
wind         -3.18137    1.81502  -1.753 0.088650 .
precip        0.51236    0.36276   1.412 0.166918
raindays     -0.05205    0.16201  -0.321 0.749972
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.64 on 34 degrees of freedom
Multiple R-squared:  0.6695,    Adjusted R-squared:  0.6112
F-statistic: 11.48 on 6 and 34 DF,  p-value: 5.419e-07
```

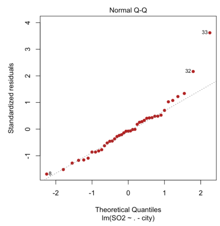
20

Third step: assess residual normality



21

Third step: assess residual normality



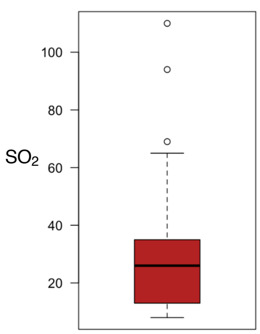
Looks suspicious

The H_0 of normality is rejected

```
> shapiro.test(residuals(fit))

        Shapiro-Wilk normality test

data:  residuals(fit)
W = 0.92303, p-value = 0.008535
```

22

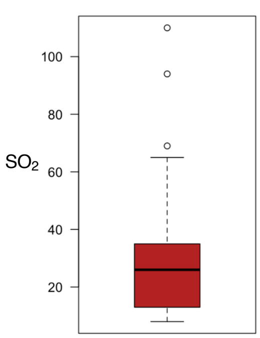
Fourth step: data transformation



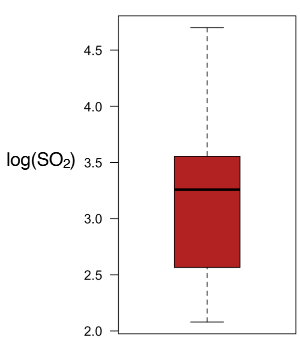
quite asymmetric

23

Fourth step: data transformation



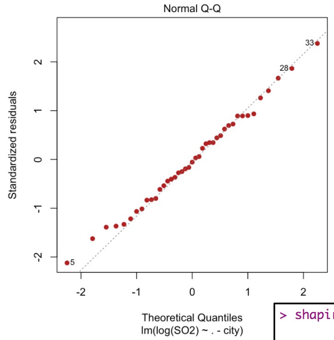
quite asymmetric



more symmetric

24

Fourth step: data transformation – re-assess normality

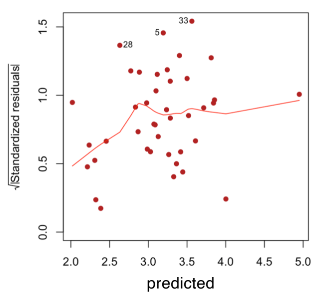


The H_0 of normality is NOT rejected after log-transformation of SO_2

```
> shapiro.test(residuals(fit.log))

        Shapiro-Wilk normality test

data:  residuals(fit.log)
W = 0.98799, p-value = 0.937
```

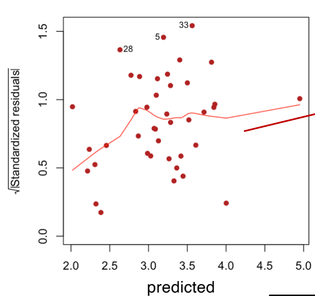
25

Fifth step: assess homoscedasticity



26

Fifth step: assess homoscedasticity



Looks suspicious

The H_0 of heteroscedasticity is NOT rejected

```
> bptest(fit.log)

        studentized Breusch-Pagan test

data:  fit.log
BP = 6.2266, df = 6, p-value = 0.3983
```

27

Sixth step: assess overall significance of the regression model

- **H_0:** The total amount of predicted variation in SO_2 is the same amount as a regression model based on the mean SO_2 (this is referred to the null model of a regression or an intercept only model).
- **H_A:** The total amount of predicted variation in SO_2 is greater than the regression model based on the average SO_2.

28

Sixth step: assess overall significance of the regression model

- H_0: The total amount of predicted variation in SO_2 is the same amount as a regression model based on the mean SO_2 (this is referred to the null model of a regression or a intercept only model).
- H_A: The total amount of predicted variation in SO_2 is greater than the regression model based on the average SO_2.

```
> summary(fit.log)

Call:
lm(formula = log(SO2) ~ . - city, data = data.pollution)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.79548 -0.25538 -0.01968  0.28328  0.98029 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.2532456  1.4483686   5.008 1.68e-05 ***
temp        -0.0599017  0.0190138  -3.150  0.00339 ** 
manu         0.0012639  0.0004820   2.622  0.01298 *  
popul       -0.0007077  0.0004632  -1.528  0.13580    
wind        -0.1697171  0.0555563  -3.055  0.00436 ** 
precip       0.0173723  0.0111036   1.565  0.12695    
raindays     0.0004347  0.0049591   0.088  0.93066    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.448 on 34 degrees of freedom
Multiple R-squared:  0.6541,    Adjusted R-squared:  0.5931 
F-statistic: 10.72 on 6 and 34 DF,  p-value: 1.126e-06
```

reject H_0

29

Sixth step: assess overall significance of the regression model

- H_0: The total amount of predicted variation in SO_2 is the same amount as a regression model based on the mean SO_2 (this is referred to the null model of a regression or a intercept only model).
- H_A: The total amount of predicted variation in SO_2 is greater than the regression model based on the average SO_2.

```
Residual standard error: 0.448 on 34 degrees of freedom
Multiple R-squared:  0.6541,    Adjusted R-squared:  0.5931
F-statistic: 10.72 on 6 and 34 DF,  p-value: 1.126e-06
```

Degrees of freedom – numerator (model) = number of predictors (p = 6)
denominator (error or residual) = n (41) – p (6) – 1 = 34

One way of reporting:

A multiple linear regression model was fit to predict SO_2 concentrations across major US cities as a function of different factors. A significant regression was found ($F_{(6,34)}$ = 10.72, P < 0.0001), with an adjusted R^2 of 0.593.

30

Seventh step: assess predictor significance

For each predictor (test the partial coefficient):

- H_0: The partial contribution of β_1 is zero.
- H_A: The partial contribution of β_1 is different from zero.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + e$$

$\beta_1, \beta_2, \dots, \beta_p$ Partial regression coefficients (or partial slopes)

31

Seventh step: assess predictor significance

For each predictor (test the partial coefficient):

H_0: The partial contribution of β_1 is zero.

H_A: The partial contribution of β_1 is different from zero.

```
> summary(fit.log)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.2532456  1.4483686   5.008 1.68e-05 ***
temp        -0.0599017  0.0190138  -3.150  0.00339 **
manu         0.0012639  0.0004820   2.622  0.01298 *
popul       -0.0007077  0.0004632  -1.528  0.13580
wind        -0.1697171  0.0555563  -3.055  0.00436 **
precip       0.0173723  0.0111036   1.565  0.12695
raindays     0.0004347  0.0049591   0.088  0.93066
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

32

Eighth step: contrast importance of predictors

Predictors are expressed in the ratio of the response unit / predictor unit; as such we can't compare their values directly.

For example, the partial slope of manufacturing is significant but its slope is much smaller than the slope of precipitation which is not significant.

As such, we need to standardize the response and predictor values (mean = 0, standard deviation = 1). As such, they will all become dimensionless (unit less) and vary in a common scale.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.2532456  1.4483686   5.008 1.68e-05 ***
temp        -0.0599017  0.0190138  -3.150  0.00339 **
manu         0.0012639  0.0004820   2.622  0.01298 *
popul       -0.0007077  0.0004632  -1.528  0.13580
wind        -0.1697171  0.0555563  -3.055  0.00436 **
precip       0.0173723  0.0111036   1.565  0.12695
raindays     0.0004347  0.0049591   0.088  0.93066
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

33

Eighth step: contrast importance of predictors

semi-partial regression coefficients

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.2532456  1.4483686   5.008 1.68e-05 ***
temp        -0.0599017  0.0190138  -3.150  0.00339 **
manu         0.0012639  0.0004820   2.622  0.01298 *
popul       -0.0007077  0.0004632  -1.528  0.13580
wind        -0.1697171  0.0555563  -3.055  0.00436 **
precip       0.0173723  0.0111036   1.565  0.12695
raindays     0.0004347  0.0049591   0.088  0.93066
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

semi-partial *standardized* regression coefficients

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.958e-16  9.962e-02   0.000  1.00000
temp        -6.165e-01  1.957e-01  -3.150  0.00339 **
manu         1.014e+00  3.868e-01   2.622  0.01298 *
popul       -5.836e-01  3.820e-01  -1.528  0.13580
wind        -3.452e-01  1.130e-01  -3.055  0.00436 **
precip       2.912e-01  1.861e-01   1.565  0.12695
raindays     1.641e-02  1.872e-01   0.088  0.93066
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

34

General linear models (not Generalized linear model)

Linear Model	Common name
✓ $Y = \mu + X$	Simple linear regression
✓ $Y = \mu + A_1$	One-factorial (one-way) ANOVA
✓ $Y = \mu + A_1 + A_2 + A_1 \times A_2$	Two-factorial (two-way) ANOVA
✓ $Y = \mu + A_1 + X\ (+A_1 \times X)$	Analysis of Covariance (ANCOVA)
✓ $Y = \mu + X_1 + X_2 + X_3$	Multiple regression
⇨ $Y = \mu + A_1 + g + A_1 \times g$	Mixed model ANOVA
$Y_1 + Y_2 = \mu + A_1 + A_2 + A_1 \times A_2$	Multivariate ANOVA (MANOVA)

Y (response) is a continuous variable
X (predictor) is a continuous variable
A represents categorical predictors (factors)
g represents groups of data (more on this later)

$(+A_1 \times X)$ - step 1 on an ANCOVA, but not in the final analysis
Multiple factors $A_1 + A_2 + \text{etc}$ (and their interactions)

35