2.8 Homework
Week 1, STAT5120, Spring 2018

1. Show that the standard deviation of *z*-scores equals 1. This can be shown as follows. First, we note that the difference in population observations minus the population mean (or in the case of samples, the sample observations minus the sample mean) are divided by the population variance σ (or in the case of samples, divided by the sample variance s). Dividing by the population/sample variances is the means of "standardizing" our normalized scores and thus rendering them within standard deviations equal to 1. Second, this idea can be demonstrated both empirically and mathematically. We see *empirically* by plotting our z-scores and if they are scores of what was originally a *normal* distribution of data, then we would see about 68.27% of our scores falling within ±1 standard devation of the population/sample mean. Chatterjee comments on the importance of standardizing our data to derive the linear correlation coefficient[1]

An example will show that the standard deviation of z-scores equals 1. We will look at two ficticious data points of x=2, x=4 at n=2:

1. Mean + sd: $\frac{2+4}{2} = 3$; $\sqrt{\frac{(2-3)^2+(4-3)^2}{2}} = \sqrt{\frac{(-1)^2+(1)^2}{2}} = \sqrt{\frac{2}{2}} = \sqrt{1} = 1$

2. Convert to z-scores

$$\frac{2-3}{1} = \frac{-1}{1} = -1$$

$$\frac{4-3}{1} = \frac{1}{1} = 1$$

3. $\mu_z = \frac{-1+1}{2} = \frac{0}{2} = 0$

$$sd_z = \sqrt{\frac{(-1)^2+(1)^2}{2}} = \sqrt{\frac{1+1}{2}} = \sqrt{\frac{2}{2}} = \sqrt{1} = 1$$

2. Refer to the original *Used Cars* example with ten observations. Our least squares regression coefficients for predicting the price of a used car is *ŷ = 199.03 + -12.15x*

(a) Use the regression line we obtained with R to predict the price of a 4.5 year-old car. Based on the above regression equation called in R with the lm() function, we could "predict" an average price of a used car by plugging a hypothesized value of 4.5 into the variable x and find that our predicted price is: 199.03 + (-12.15*4.5) = $144.355

(b) Use the regression line we obtained with R to predict the price of a 60 year-old car. Does this prediction make any sense? Explain. What went wrong? Using our regression equation given above, we would obtain a predicted price of ($529.97) but this predicted negative price is problematic for two reasons. One, there aren't too many negatively priced used cars available (if there are, I'd like to cash in on them) and two, for purposes of getting this answer correct, our regression equation was calculated based on a specific *range* of car ages from 3 to 9 years in age. But in this case, the age of a 60-year old car *exceeds* the useful (and allowable) range upon which our regression equation was built.

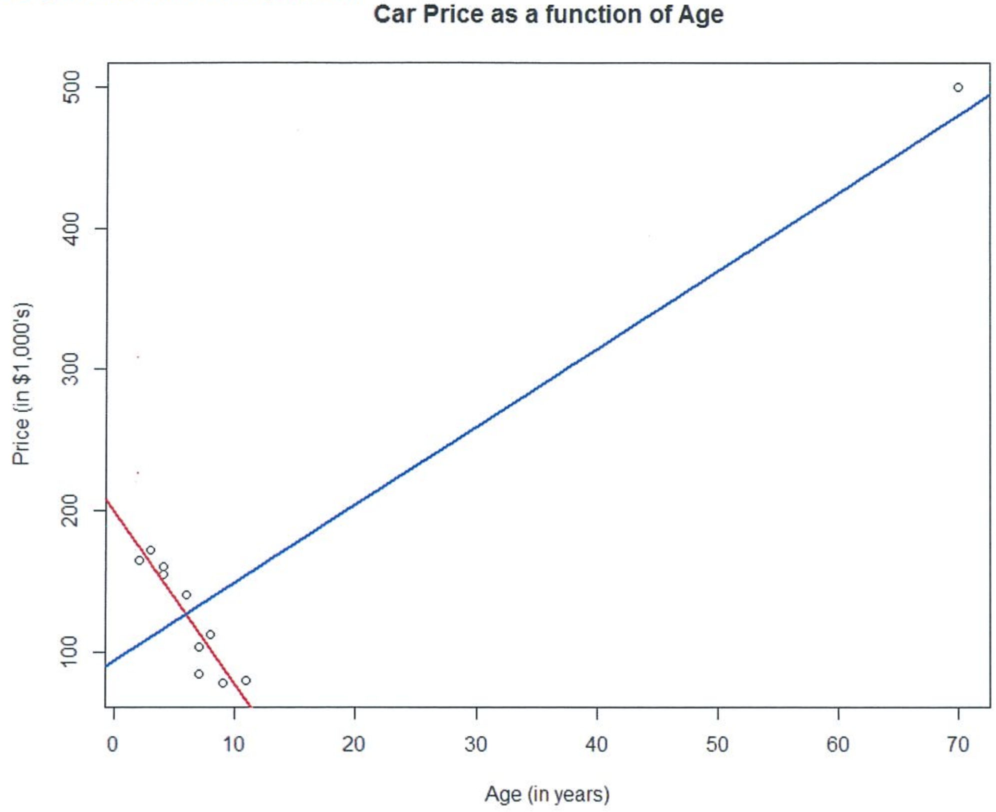
(c) Suppose your friend just sold a 70 year-old car for $50,000, and so we add that to our data (70, 500). You can use the rbind() command.

i. What is the new regression line equation? *ŷ = 93.226 + 5.523x*

[1] Chatterjee, Samprit, and Ali S. Hadi, *Regression Analysis by Example*, 5th ed., (Hoboken: John Wiley & Sons, 2012), 27-28.

ii. What is the new *r* value? Our new r^2 for this equation is now 0.8209 and if we derive the square root of this, we find that our new *r* value is 0.9060353

iii. Make a scatterplot of the new data, and include both regression lines (original and new one). Is the new regression line equation better or worse than the original one for predicting used car prices? Explain. The scatter plot below shows the new bivariate pattern after adding a 70-year old car purchase of $50,000. The r2 is high and the new regression line (blue) goes through the data in some way, the bivariate pattern shows that the new line, and hence the equation itself, is really useless. The new data point is an influential point that is perpendicular to the downward slope of all other used care purchases which are shown their regression line in red. The new regression results, and their corresponding added data point, are therefore worse than the original set with its equation.

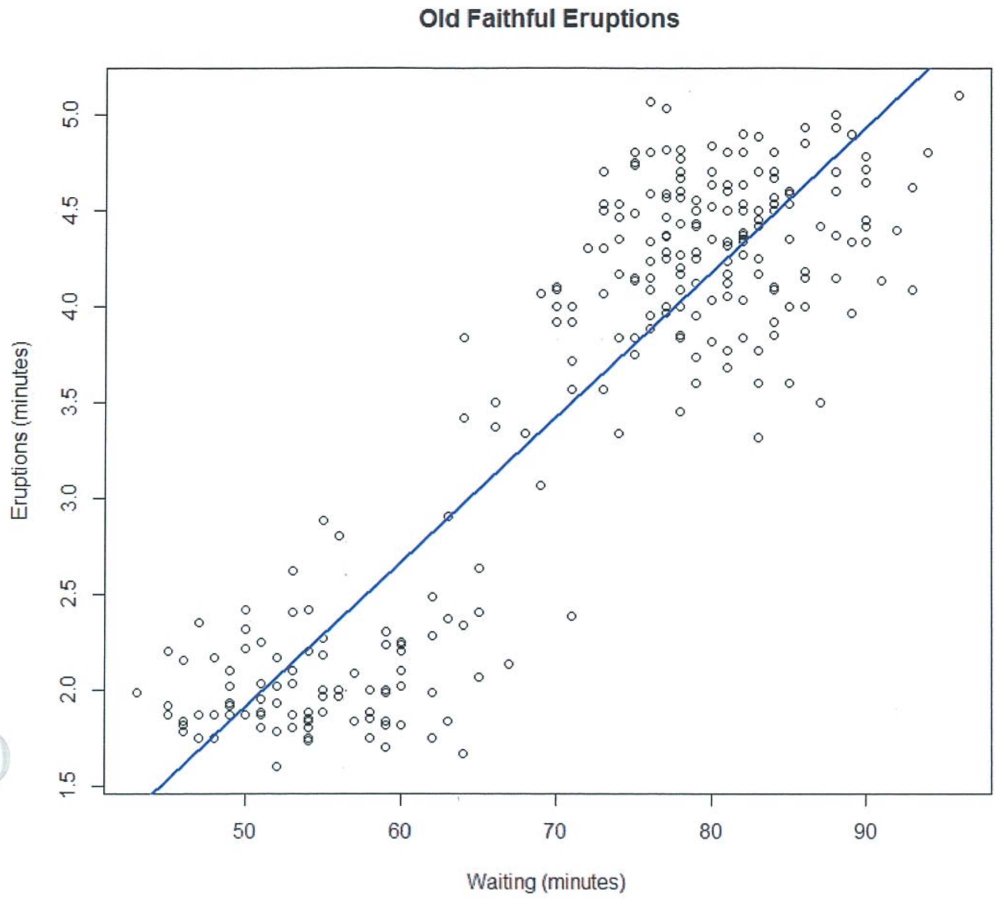


iv. How is this new observation affecting our model? This new observation has the affect of "pulling" our new regression line (blue) to become perpendicular to the other data points compared to the original model.

v. Remark on the fact that the new coefficient of determination is reasonably high, yet the new model is not very good. The new coefficient of determination is reasonably high yet the model is not very useful since the new model assumes the regression results (and line in blue) should be perpendicular to all other original points but an examination of the scatter plot above shows this hypothesis to be unreasonable.

3. Open the data set called *faithful* in R simply by typing *faithful* at the prompt. The first column contains eruption durations in minutes and the second contains the times between eruptions, also in minutes. The data are rounded quite a bit- a minute is a relatively long time. Use R to make a scatterplot of the duration vs. the times between. Also perform linear regression. Remark on the plot and the regression output, especially on r^2 and how it corresponds to the point pattern in the plot. A scatterplot was constructed and shown below. A linear regression was also calculated to demonstrate a possible linear relationship between Eruptions (in minutes) regressed on Waiting (in minutes), yielding the following equation: $\hat{y} = -1.874016 + 0.075628x$

The r2 is 0.8115 which is reasonably high. This result seems to correspond nicely with the estimated least-squares line fitting squarely down the center of the data points. Interesting, there appear to be two general clusters in the bivarite plot below, though without further research, I cannot confirm whether this is merely an "artifact" of the data itself or if there are additional geophysical forces, and hence unseen factors, at work.



4. Show the slope coefficient is indeed $b_1 = r * s_y/s_x$.

$$\frac{d}{db_1}\sum_{i=1}^{n} e_i^2 = -2\sum_{i=1}^{n}(y_i - b_i x_i - b_0)x_i$$

$$\varnothing = -2\sum_{i=1}^{n}(y_i - b_i x_i - (\bar{y} - b_1\bar{x})x_i$$

5. Explain how the ratio s_y/s_x affects the slope of the regression line through the original data. The slope of the regression equation is given in question 4 as being $b_1 = r * s_y/s_x$. Now, in examining this equation, we see that the ratio of the S_y over the S_x leads us to conclude that mathematically as the S_y numerator becomes greater and greater than the denominator, the importance of the y variable becomes greater. The regression line then becomes "steeper," resembling perhaps the long-shot odds of the Miami Dolphins ever getting into the playoffs before the end of the 21st Century. On the other hand, as the S_y numerator becomes less than the denominator, the regression line tends to flatten out. In short, this ratio shows the "weight" of either the y or x variables in the regression slope, whether it will be steeper or flatter.

6. Show that the regression line that runs through the standardized data values has the form $\hat{z}_y = r * z_x$

Our standardized data points have the same bivariate relationships, albeit re-scaled to have $\mu = 0$ + $sd = 1$. The regression line runs through point (0, 0) as a scatter plot shows.

Standardizing removes the unit size effect.

7. Show that the cofficient of determination is indeed equal to the square of the linear correlation coefficient (hence the nickname "r^2").

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$= 1 - \frac{\Sigma(y_i - \hat{y})^2}{\Sigma(y_i - \bar{y})^2}$$

$$= \frac{Cov(y, \hat{y})}{\sqrt{Var(y)\,Var(\hat{y})}}$$

$$\rightarrow \left(\frac{Cov(y, \hat{y})}{\sqrt{Var(y)\,Var(\hat{y})}}\right)^2$$

$$= \frac{Cov(y, \hat{y})\,Cov(y, \hat{y})}{Var(y)\,Var(\hat{y})}$$

$$= \frac{Var(\hat{y})\,Var(\hat{y})}{Var(y)\,Var(\hat{y})}$$

$$= \frac{Var(\hat{y})}{Var(y)} = \frac{SSE}{SST} = R^2$$

8. Show that for any point estimator θ̂ that MSE(θ̂) = B(θ̂)² + Var(θ̂)

$$MSE_\theta = E(\hat{\theta} - \theta)^2 = Var(\hat{\theta}) + (E(\theta) - \theta)^2 = Var(\hat{\theta}) + (\text{Bias of } \hat{\theta})^2$$

This is so because:

$$E(\hat{\theta} - \theta)^2 = E(\hat{\theta}^2) + E(\theta^2) - 2\theta E(\hat{\theta})$$

$$= Var(\hat{\theta}) + [E(\theta)]^2 + \theta^2 - 2\theta E(\hat{\theta})$$

$$= Var(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2$$

9. Prove that the sample variance is an unbiased estimator for the population variance. That is, show that

$$E\left[\frac{1}{n-1}\sum_{k=1}^{n}(Y_i - \overline{Y})^2\right] = \sigma^2.$$

$$(n-1)\,E[\sigma^2] = E\left[\sum_{i=1}^{n} x_i^2 - 2\bar{x}x_i + \bar{x}^2\right]$$

$$\frac{n-1}{n}\,E[\sigma^2] = E[x_i^2] - E[\bar{x}^2]$$

$$\frac{n-1}{n}\,E[\sigma^2] = [\sigma^2 + \mu^2] - [\tfrac{1}{n}\sigma^2 + \mu]$$

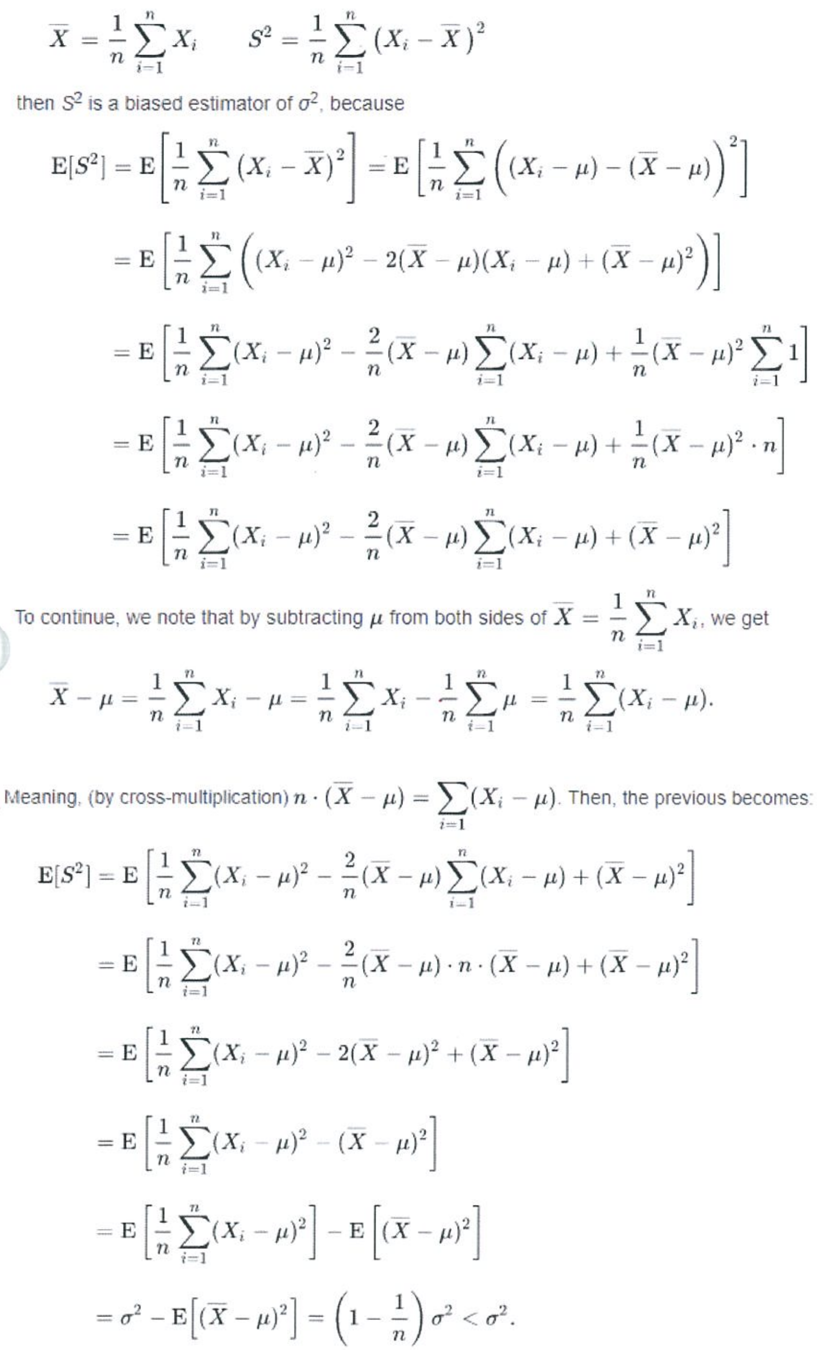
$$\frac{n-1}{n}\,E[\sigma^2] = \sigma^2 - \tfrac{1}{n}\sigma^2 \longrightarrow E[\sigma^2] = \sigma^2$$

10. Verify the bias of

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$$

in estimating the population variance is *–σ2/n*

The following source demonstrates this proof:[2]



$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i \qquad S^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\,\right)^2$$

then S^2 is a biased estimator of σ^2, because

$$\begin{aligned}
\mathrm{E}[S^2] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\bigg((X_i-\mu)-(\overline{X}-\mu)\bigg)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\bigg((X_i-\mu)^2-2(\overline{X}-\mu)(X_i-\mu)+(\overline{X}-\mu)^2\bigg)\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)+\frac{1}{n}(\overline{X}-\mu)^2\sum_{i=1}^{n}1\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)+\frac{1}{n}(\overline{X}-\mu)^2\cdot n\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)+(\overline{X}-\mu)^2\right]
\end{aligned}$$

To continue, we note that by subtracting μ from both sides of $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$, we get

$$\overline{X}-\mu=\frac{1}{n}\sum_{i=1}^{n}X_i-\mu=\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}\mu \;=\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu).$$

Meaning, (by cross-multiplication) $n\cdot(\overline{X}-\mu)=\sum_{i=1}(X_i-\mu)$. Then, the previous becomes:

$$\begin{aligned}
\mathrm{E}[S^2] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\frac{2}{n}(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)+(\overline{X}-\mu)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\frac{2}{n}(\overline{X}-\mu)\cdot n\cdot(\overline{X}-\mu)+(\overline{X}-\mu)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-2(\overline{X}-\mu)^2+(\overline{X}-\mu)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-(\overline{X}-\mu)^2\right] \\
&= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right]-\mathrm{E}\left[(\overline{X}-\mu)^2\right] \\
&= \sigma^2-\mathrm{E}\left[(\overline{X}-\mu)^2\right]=\left(1-\frac{1}{n}\right)\sigma^2<\sigma^2.
\end{aligned}$$

[2] "Bias of an Estimator," Wikipedia, accessed at: https://en.wikipedia.org/wiki/Bias_of_an_estimator

APPENDIX: R SCRIPTS USED

Question 2(c):

```
> age <- c(3,6,8,4,2,11,4,7,7,9)
> price <- c(172,140,112,160,165,80,155,103,84,78)
> used_cars <- data.frame(age,price)
> used_cars_reg <- lm(used_cars$price ~ used_cars$age)
> more_used_cars <- rbind(used_cars, c(70,500))
> more_used_cars
   age price
1    3   172
2    6   140
3    8   112
4    4   160
5    2   165
6   11    80
7    4   155
8    7   103
9    7    84
10   9    78
11  70   500
> more_used_cars_reg <- lm(more_used_cars$price ~ more_used_cars$age)
> summary(more_used_cars_reg)

Call:
lm(formula = more_used_cars$price ~ more_used_cars$age)

Residuals:
   Min     1Q Median     3Q    Max
-73.98 -38.39  13.64  42.18  62.20

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         93.2263    18.9534   4.919 0.000826 ***
more_used_cars$age   5.5230     0.8598   6.423 0.000122 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 52.9 on 9 degrees of freedom
Multiple R-squared:  0.8209,	Adjusted R-squared:  0.801
F-statistic: 41.26 on 1 and 9 DF,  p-value: 0.0001219

> plot(more_used_cars$price ~ more_used_cars$age, main="Car Price as a function of Age", xlab="Age (in years)",
ylab="Price (in $1,000's)")
> abline(used_cars_reg, lwd=2, col="red")
> abline(more_used_cars_reg, lwd=2, col="blue")
```

Question 3:

```
> faithful
> summary(faithful_reg <- lm(faithful$eruptions ~ faithful$waiting))
```

```
Call:
lm(formula = faithful$eruptions ~ faithful$waiting)

Residuals:
     Min       1Q   Median       3Q      Max
-1.29917 -0.37689  0.03508  0.34909  1.19329

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -1.874016   0.160143  -11.70   <2e-16 ***
faithful$waiting  0.075628   0.002219   34.09   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4965 on 270 degrees of freedom
Multiple R-squared:  0.8115,	Adjusted R-squared:  0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16

> plot(faithful$eruptions ~ faithful$waiting, main="Old Faithful Eruptions", xlab="Waiting (minutes)", ylab="Eruptions
(minutes)")
> abline(faithful_reg,lwd=2, col="blue")
```