# PSTAT 126 - Assignment 2

Fall 2022

Kayla Katakis

*Note:* ***Submit both your `Rmd` and generated pdf file to Canvas.*** *Use the same indentation level as* ***Solution*** *markers to write your solutions. Improper indentation will break your document.*

```
library(alr4)
library(ggplot2)
data(UN11)
```

1. The data set `UN11` in the `alr4` package contains several variables, including `ppgdp`, the gross national product per person in U.S. dollars, and `fertility`, the birth rate per 1000 females, from the year 2009. The data are for 199 localities, and we will study the regression of `fertility` on `ppgdp`.

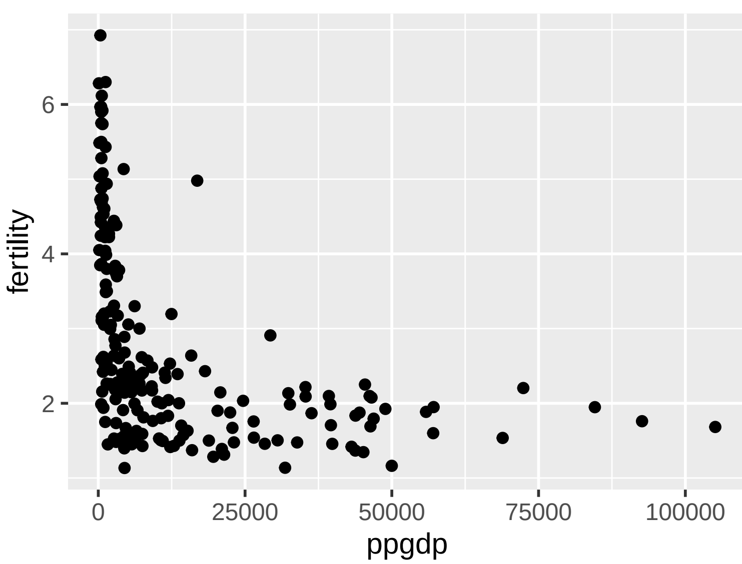
(a) Identify the predictor and response.

**Solution**: The predictor is ppgdp, and the response is fertility.

(b) Draw the scatterplot of fertility against ppgdp and describe the relationship between these two variables. Is the trend linear?

**Solution**: The trend is not linear, it appears to be negatively exponential. As fertility increases, ppgdp decreases dramatically.

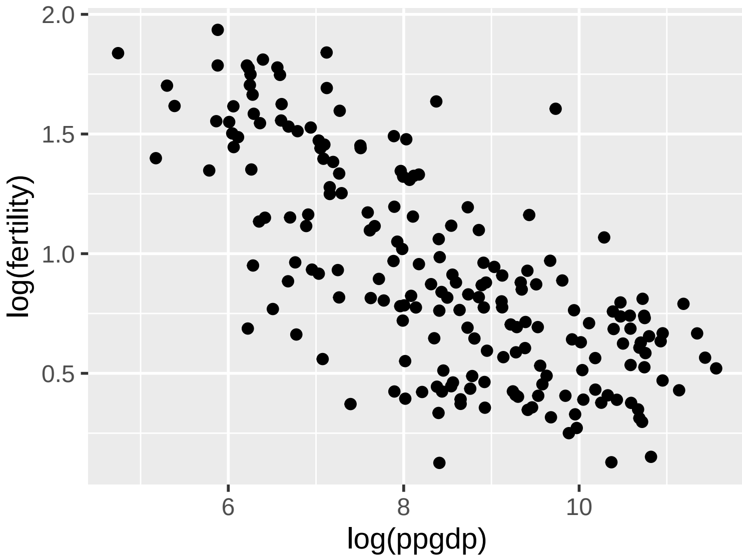
```
ggplot(UN11, aes(ppgdp, fertility)) + geom_point()
```




(c) Replace both variables by their natural logarithms and draw another scatterplot. Does the simple linear regression model seem plausible for a summary of this graph?

**Solution**: The simple linear regression model definitely seems more plausible, as taking the natural log of each variable resulyts in a more linear relationship with a clearer slope and intercept.

```
ggplot(UN11, aes(log(ppgdp),log(fertility))) + geom_point()
```

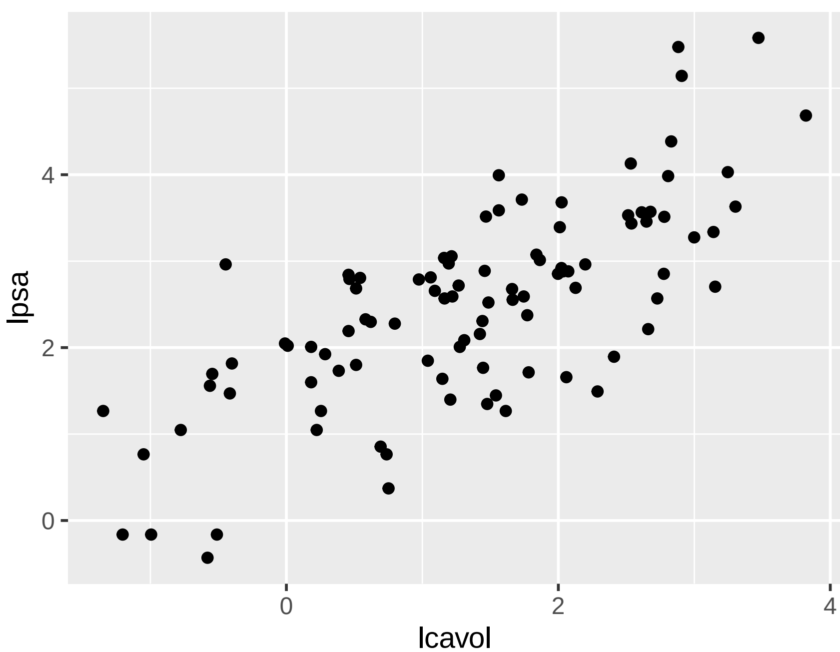



2. The data set `prostate` in the `faraway` package is from a study of 97 men with prostate cancer. Interest is in predicting `lpsa` (log prostate specific antigen) with `lcavol` (log cancer volume). You may not use the function `lm` for this question.

(a) Draw a scatterplot - does a simple linear regression model seem reasonable?

**Solution**: A simple linear regression model seems reasonable; there is a clear positive linear relationship

```
data("prostate", package = "faraway")
View(prostate)
ggplot(prostate, aes(lcavol, lpsa)) + geom_point()
```



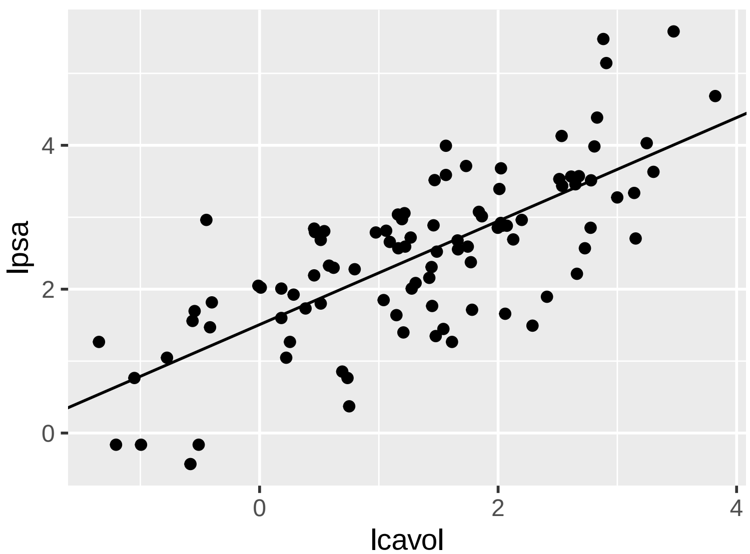
(b) Compute the values $\bar{x}, \bar{y}, S_{xx}, S_{yy}$ and $S_{xy}$. Compute the ordinary least squares estimates of the intercept and slope for the simple linear regression model, and draw the fitted line on your plot from part a).

**Solution:**

```
xbar = mean(prostate$lcavol)
ybar = mean(prostate$lpsa)
sxx = sum((prostate$lcavol - xbar)**2)
syy = sum((prostate$lpsa - ybar)**2)
sxy = sum((prostate$lcavol - xbar)*(prostate$lpsa - ybar))

OLS_b1 = sxy/sxx
OLS_b0 = ybar - OLS_b1*xbar

ggplot(prostate, aes(lcavol, lpsa)) + geom_point()+ geom_abline(aes(intercept=OLS_b0, slope = OLS_b1))
```




(c) Compute $\hat{\sigma}^2$ and find the estimated standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$. Also find the estimated covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$.

**Solution:** Note that $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$

```
error = prostate$lpsa - (OLS_b0 +OLS_b1*prostate$lcavol)
RSS = sum(error**2)
sigma2hat = (1/nrow(prostate)-2)*RSS

se_b0 = sigma2hat*((1/nrow(prostate))+(xbar**2/sxx))
se_b1 = sigma2hat/sxx
covb0b1 = ((-xbar*sigma2hat)/sxx)
```

(d) Carry out $t$-tests for the two null hypotheses $\beta_0 = 0$ and $\beta_1 = 0$, reporting the value of the test statistic and a $p$-value in each case.

**Solution:** In both cases, where $\beta_0 = 0$ and $\beta_1 = 0$, as seen in the model, the p-values are the same at 2.2e-16.

```
model_prostate <- lm(lpsa~lcavol, prostate)
summary(model_prostate)
```

```
##
## Call:
```

```
## lm(formula = lpsa ~ lcavol, data = prostate)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67625 -0.41648  0.09859  0.50709  1.89673
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.50730    0.12194   12.36   <2e-16 ***
## lcavol       0.71932    0.06819   10.55   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7875 on 95 degrees of freedom
## Multiple R-squared:  0.5394, Adjusted R-squared:  0.5346
## F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16
```

3. The data set `ftcollinstemp` in the `alr4` package gives the mean temperature in the fall of each year, defined as September 1 to November 30 , and the mean temperature in the following winter, defined as December 1 to the end of February in the following calendar year, in degrees Fahrenheit, for Ft. Collins, CO (Colorado Climate Center, 2012). These data cover the time period from 1900 to 2010 . The question of interest is: Does the average fall temperature predict the average winter temperature?

(a) Use the `lm` function in R to fit the regression of the response on the predictor. Draw a scatterplot of the data and add your fitted regression line.

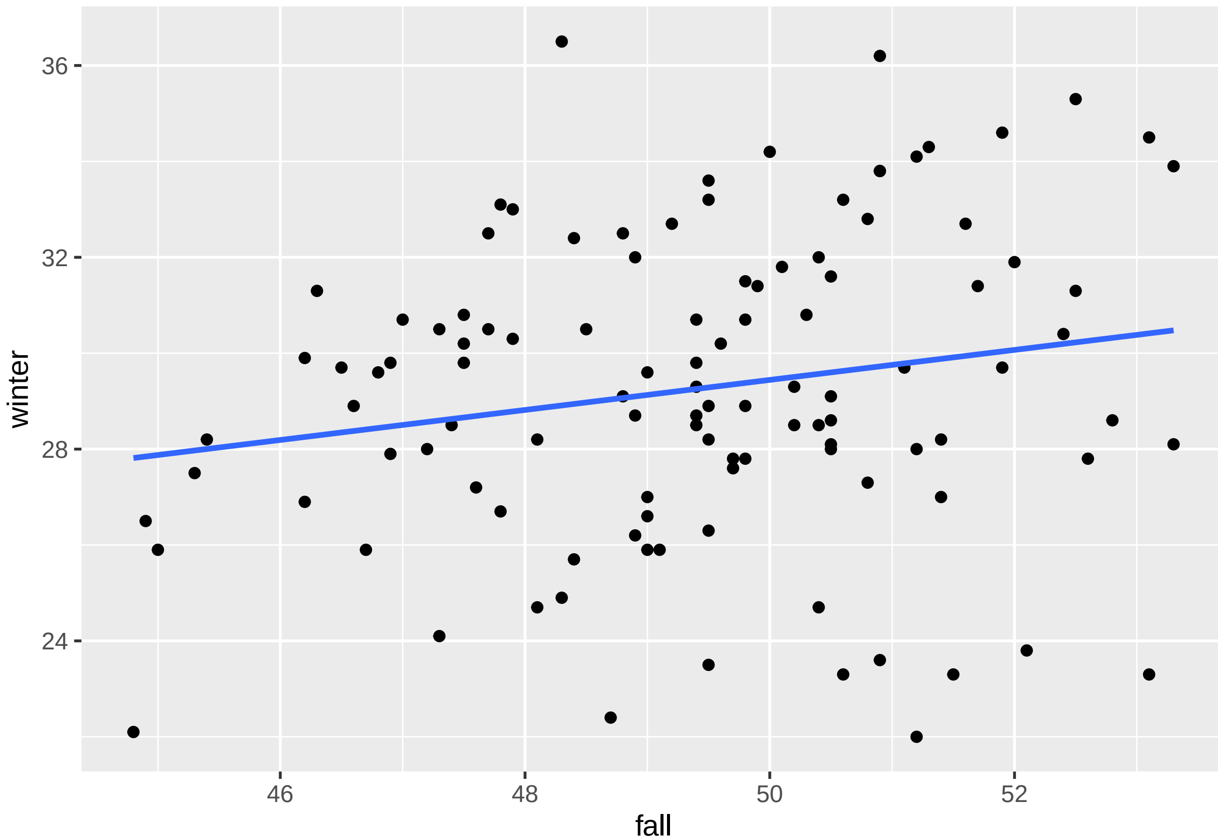
**Solution:**

```r
data("ftcollinstemp", package = 'alr4')
temp_lm <- lm(winter ~ fall, ftcollinstemp)
summary(temp_lm)
```

```
##
## Call:
## lm(formula = winter ~ fall, data = ftcollinstemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8186 -1.7837 -0.0873  2.1300  7.5896
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.7843     7.5549   1.825   0.0708 .
## fall          0.3132     0.1528   2.049   0.0428 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.179 on 109 degrees of freedom
## Multiple R-squared:  0.0371, Adjusted R-squared:  0.02826
## F-statistic:   4.2 on 1 and 109 DF,  p-value: 0.04284
```

```r
ggplot(ftcollinstemp, aes(fall, winter)) + geom_point() + stat_smooth(method = 'lm', se = FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




(b) Test the null hypothesis that the slope is 0 against a two-sided alternative at $\alpha = 0.01$, and interpret your findings.

**Solution:** The $p$-value corresponding to $\beta_1$, or the change in rate of winter temperatures as fall temperatures increase, is 0.0428. At $\alpha = 0.01$, this value is not significant, so we fail reject the null hypothesis that the slope is 0, which implies that fall temperatures are not a significant predictor of winter temperatures.

(c) What percentage of the variability in winter is explained by fall?

**Solution:** The $R^2$ value given by the model is 0.0371, which implies that approximately 3.7% of the variability in winter temperatures is explained by fall temperatures.