

HW08

Luke Beebe

2024-04-17

```
data("water", package="alr4")
dim(water)
```

```
## [1] 43 8
```

```
head(water)
```

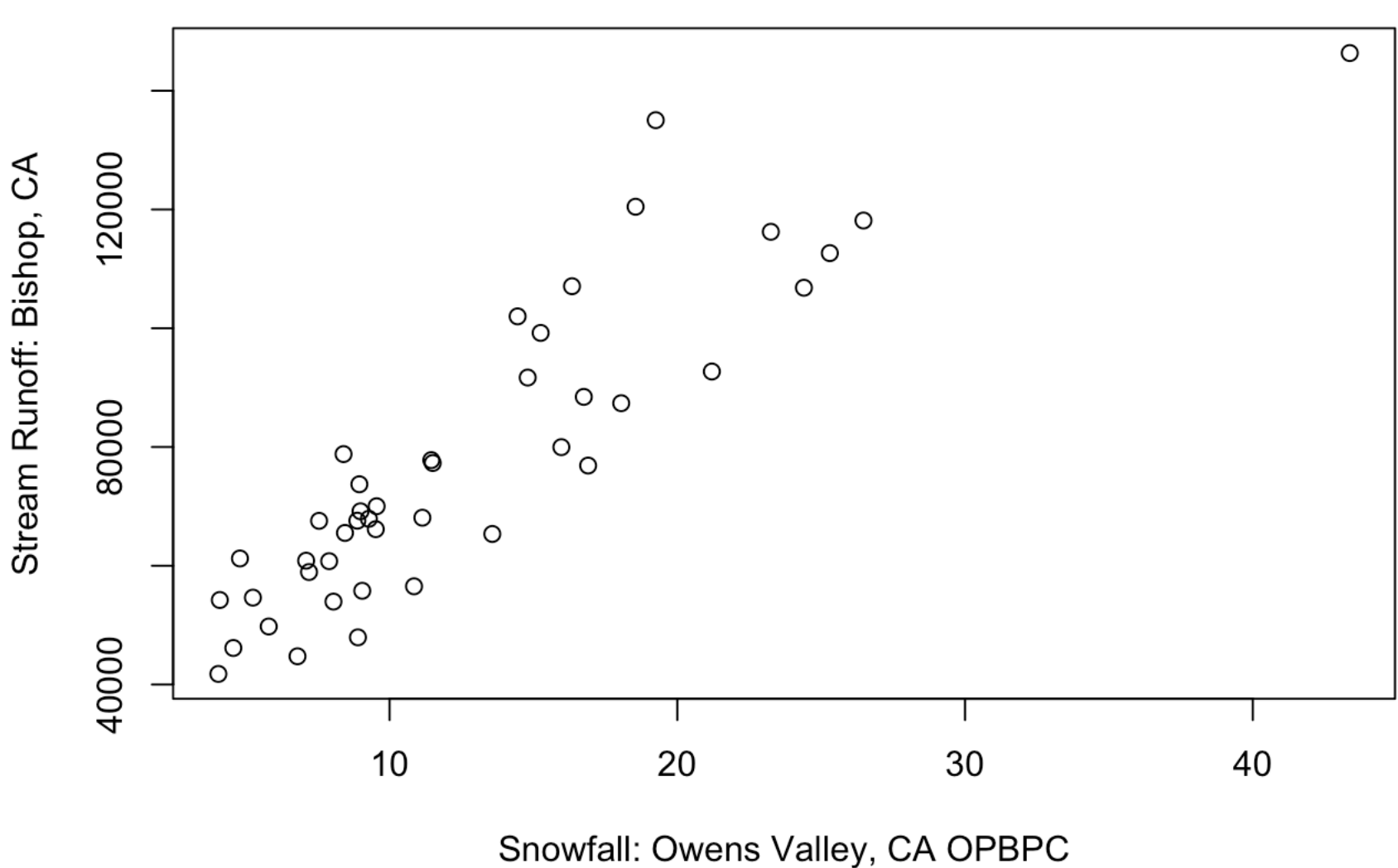
```
##      Year APMAM APSAB APSLAKE OPBPC  OPRC OPSLAKE  BSAAM
## 1 1948   9.13   3.58   3.91   4.10   7.43   6.47  54235
## 2 1949   5.28   4.82   5.20   7.55  11.11  10.26  67567
## 3 1950   4.20   3.77   3.67   9.52  12.20  11.35  66161
## 4 1951   4.60   4.46   3.93  11.14  15.15  11.13  68094
## 5 1952   7.15   4.99   4.88  16.34  20.05  22.81 107080
## 6 1953   9.70   5.65   4.91   8.88   8.15   7.41  67594
```

1

Create scatterplots of stream runoff (BSAAM) against each of the three snowfall totals (OPBPC, OPRC, OPSLAKE), making sure to carefully label all axes. Include the scatterplots in your report.

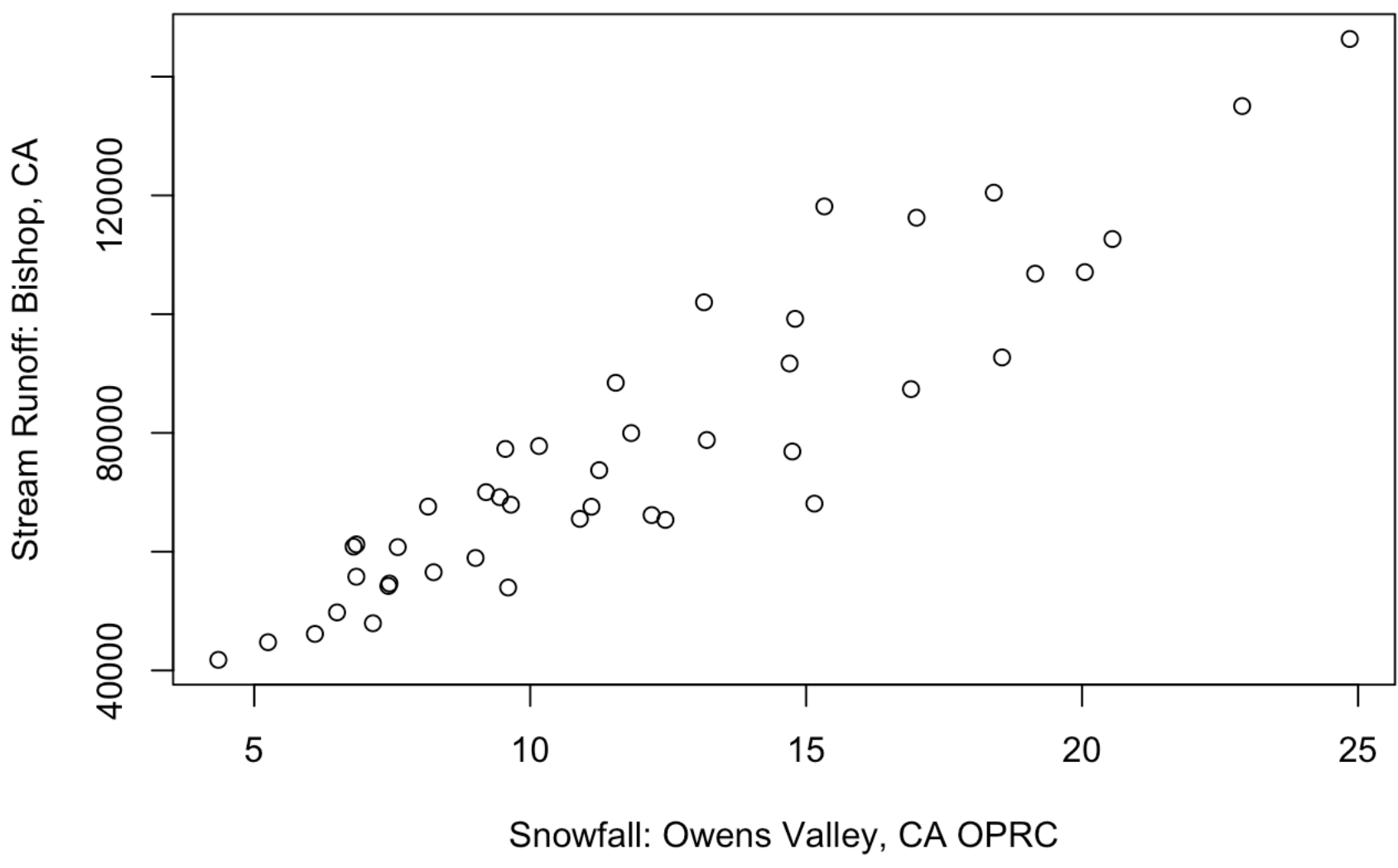
```
plot(water$OPBPC, water$BSAAM, xlab="Snowfall: Owens Valley, CA OPBPC", ylab="Stream Runoff: Bishop, CA", main="Snowfall (inches) vs Stream (acre/ft)")
```

Snowfall (inches) vs Stream (acre/ft)



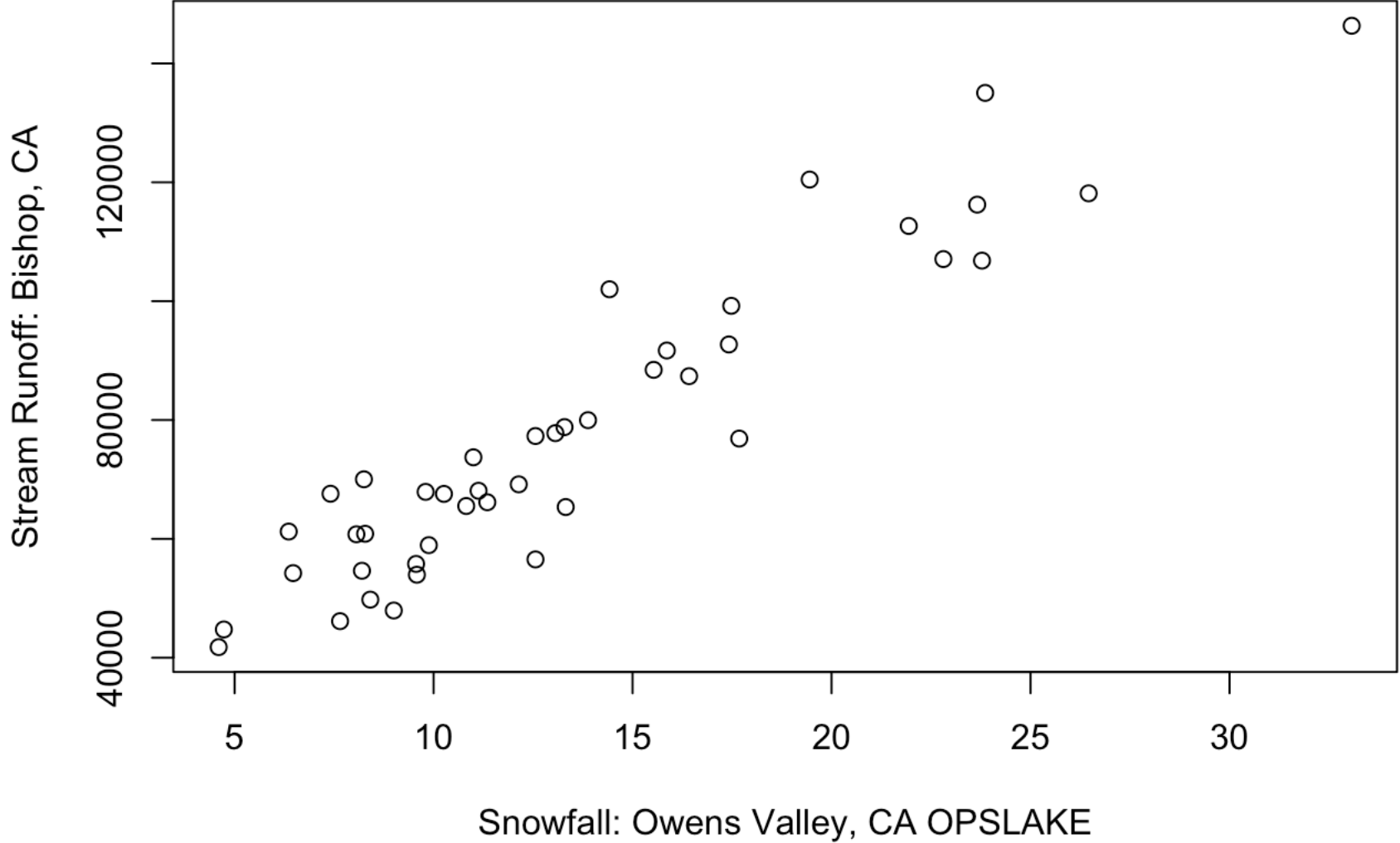
```
plot(water$OPRC, water$BSAAM, xlab="Snowfall: Owens Valley, CA OPRC", ylab="Stream Runoff: Bishop, CA", main="Snowfall (inches) vs Stream (acre/ft)")
```

Snowfall (inches) vs Stream (acre/ft)



```
plot(water$OPSLAKE, water$BSAAM, xlab="Snowfall: Owens Valley, CA OPSLAKE", ylab="Stream Runoff: Bishop, CA", main="Snowfall (inches) vs Stream (acre/ft)")
```

Snowfall (inches) vs Stream (acre/ft)



2

Write down a statistical model in which the expected stream runoff is a linear function (with intercept) of the three snowfall totals, with errors that are independent, constant variance, and normally distributed. Use the mathematical notation used in the lectures, and define all symbols and subscripts.

$BSAAM = \beta_0 + \beta_1 \cdot OPBPC + \beta_2 \cdot OPRC + \beta_3 \cdot OPSLAKE + \text{error}$
 $\text{error} \sim N(0, \sigma^2)$
 β_0 = intercept
 β_1 = coefficient for snowfall total at OPBPC
 β_2 = coefficient for snowfall total at OPRC
 β_3 = coefficient for snowfall total at OPSLAKE
 error = error at instance i
 $\sim N(0, \sigma^2)$
 $BSAAM_i$ = stream runoff at instance i
 $OPBPC_i$ = snowfall total at instance i for OPBPC location
 $OPRC_i$ = snowfall total at instance i for OPRC location
 $OPSLAKE_i$ = snowfall total at instance i for OPSLAKE location

3

Compute the least squares estimates of the regression coefficients and compute the unbiased estimate of the variance parameter. Report the answers in terms of the notation used in part (b).

```
model <- lm(BSAAM ~ OPBPC+OPRC+OPSLAKE, data=water)
summary(model)
```

```
##
## Call:
## lm(formula = BSAAM ~ OPBPC + OPRC + OPSLAKE, data = water)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15964.1  -6491.8   -404.4    4741.9  19921.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22991.85   3545.32   6.485  1.1e-07 ***
## OPBPC         40.61     502.40   0.081  0.93599
## OPRC        1867.46     647.04   2.886  0.00633 **
## OPSLAKE      2353.96     771.71   3.050  0.00410 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8304 on 39 degrees of freedom
## Multiple R-squared:  0.9017, Adjusted R-squared:  0.8941
## F-statistic: 119.2 on 3 and 39 DF,  p-value: < 2.2e-16
```

$\beta_0 = 22991.85$, highly significant, standard error = 3545.32
 $\beta_1 = 40.61$, not significant, standard error = 502.40 (95% CI contains 0)
 $\beta_2 = 1867.46$, highly significant, standard error = 647.04
 $\beta_3 = 2353.96$, highly significant, standard error = 771.71
 $\text{error} \sim N(0, 8304^2)$

$BSAAM = 22991.85 + 40.61 \cdot OPBPC + 1867.46 \cdot OPRC + 2353.96 \cdot OPSLAKE + N(0, 8304^2)$

4

Test the hypotheses that each of the regression coefficients in the multiple linear model are zero, and report the t -statistics and the p -values. What can we conclude from these tests?

beta0: t=6.485, p=1.1e-07
beta1: t=0.081, p=0.93599
beta2: t=2.886, p=0.00633
beta3: t=3.050, p=0.00410

We can conclude that **beta0**, **beta2**, and **beta3** are all significant variables in explaining the response.

5

Report the R^2 value and give its interpretation.

The R^2 of this model is 0.9017, and means that 90.17% of the variance in the model is explained by OPBPC, OPRC, and OPSLAKE.

6

You should have failed to reject one of the null hypotheses in part (d). Re-run the regression without that variable. Do the standard errors for the other coefficients get larger or smaller?

```
model <- lm(BSAAM ~ OPRC+OPSLAKE, data=water)
summary(model)
```

```
##
## Call:
## lm(formula = BSAAM ~ OPRC + OPSLAKE, data = water)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15991.2  -6484.6   -498.3    4700.1  19945.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22891.2    3277.8    6.984 1.98e-08 ***
## OPRC         1866.8     638.8    2.922  0.0057 **
## OPSLAKE      2400.5     503.3    4.770  2.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8201 on 40 degrees of freedom
## Multiple R-squared:  0.9017, Adjusted R-squared:  0.8967
## F-statistic: 183.4 on 2 and 40 DF,  p-value: < 2.2e-16
```

The standard errors for the other coefficients shrink when removing OPBPC from the model. The standard error for the model also shrinks.

7

Summarize in words your conclusions about the relationship between stream runoff and the snowfall measurements, and the implications for predicting stream runoff from snowfall.

There is a strong relationship using snowfall measurements in the locations of OPRC and OPSLAKE to explain 90.17% of the variance in stream runoff in Bishop, CA. Without repeating more of the numbers above, I can say that with this data, **our model is a good predictor and can be used to forecast stream runoff.**