testexercise1.Rmd

Ed Goodwin

May 21, 2016

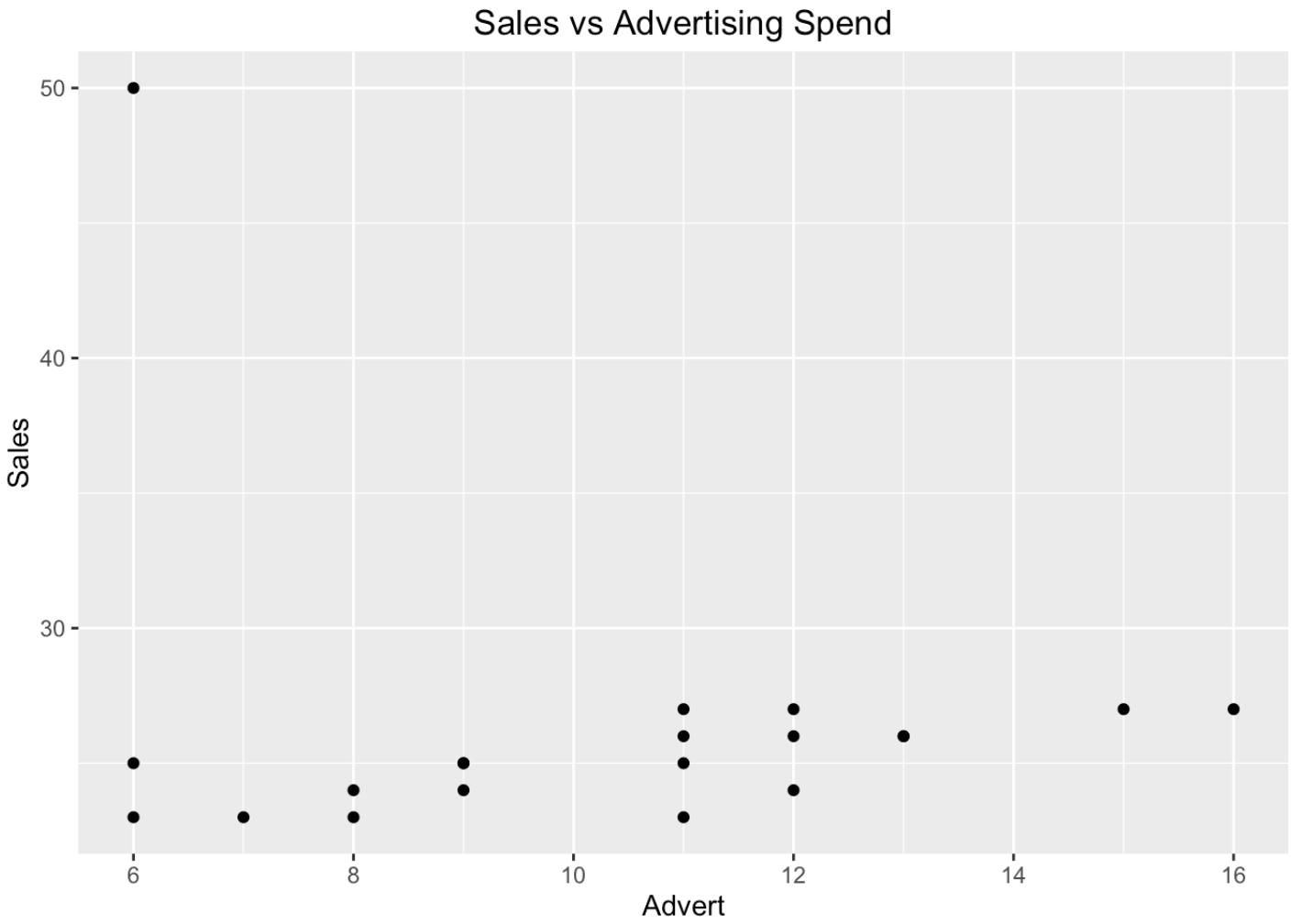
Test Exercise 1

Questions

This exercise considers an example of data that do not satisfy all the standard assumptions of simple regression.

In the considered case, one particular observation lies far off from the others, that is, it is an outlier. This violates assumptions A3 and A4, which state that all error terms are drawn from one and the same distribution with mean zero and fixed variance. The dataset contains twenty weekly observations on sales and advertising of a department store. The question of interest lies in estimating the effect of advertising on sales. One of the weeks was special, as the store was also open in the evenings during this week, but this aspect will first be ignored in the analysis.

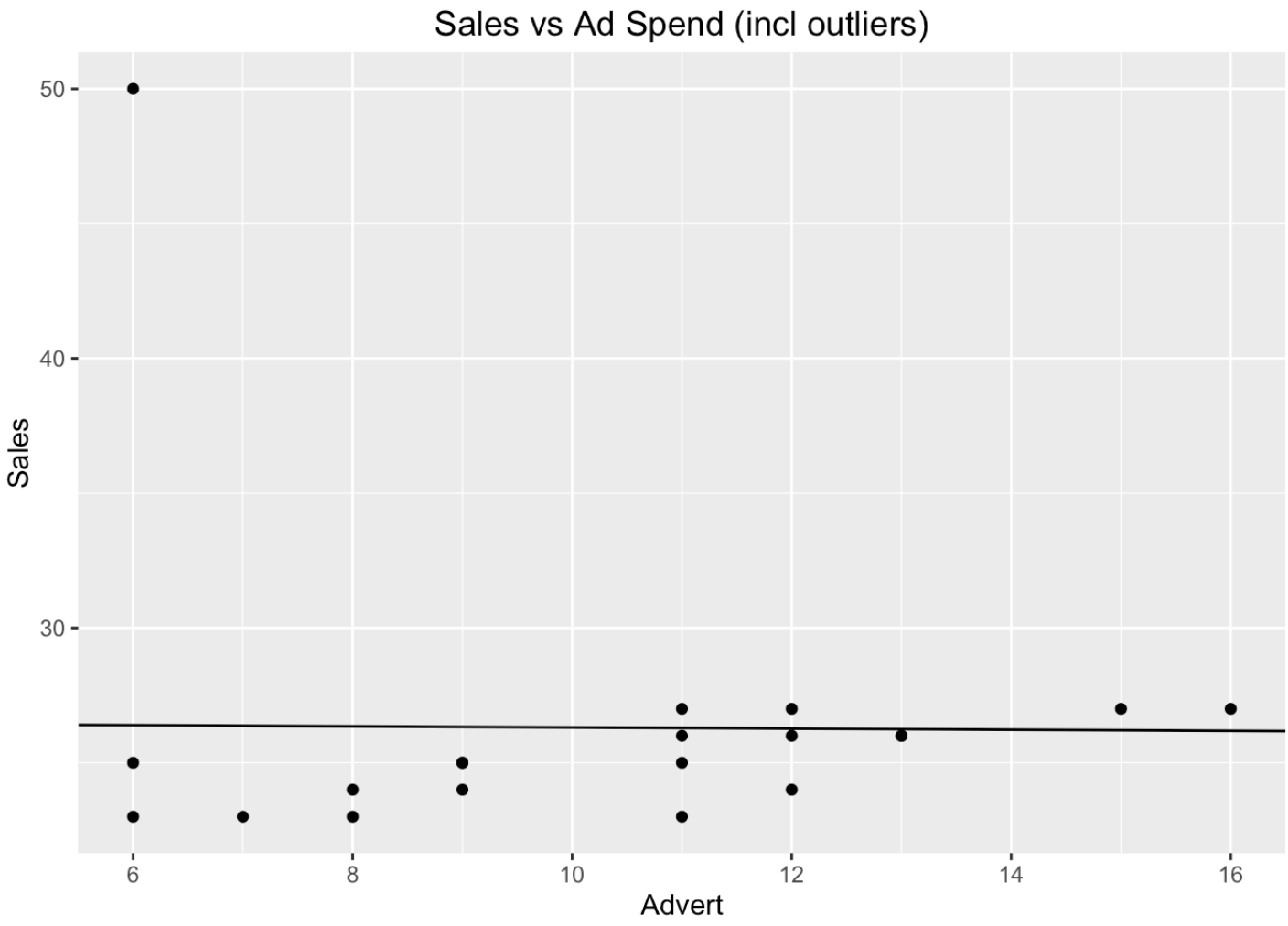
Make the scatter diagram with sales on the vertical axis and advertising on the horizontal axis.



What do you expect to find if you would fit a regression line to these data?

The regression line would be misleading because of the outlier in Observation 12.

```
##
## Call:
## lm(formula = tsalesdat$Sales ~ tsalesdat$Observ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4158 -2.2316 -1.3211  0.0395 23.7316
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      26.52105    2.74800   9.651 1.54e-08 ***
## tsalesdat$Observ -0.02105    0.22940  -0.092    0.928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.916 on 18 degrees of freedom
## Multiple R-squared:  0.0004677,  Adjusted R-squared:  -0.05506
## F-statistic: 0.008422 on 1 and 18 DF,  p-value: 0.9279
```

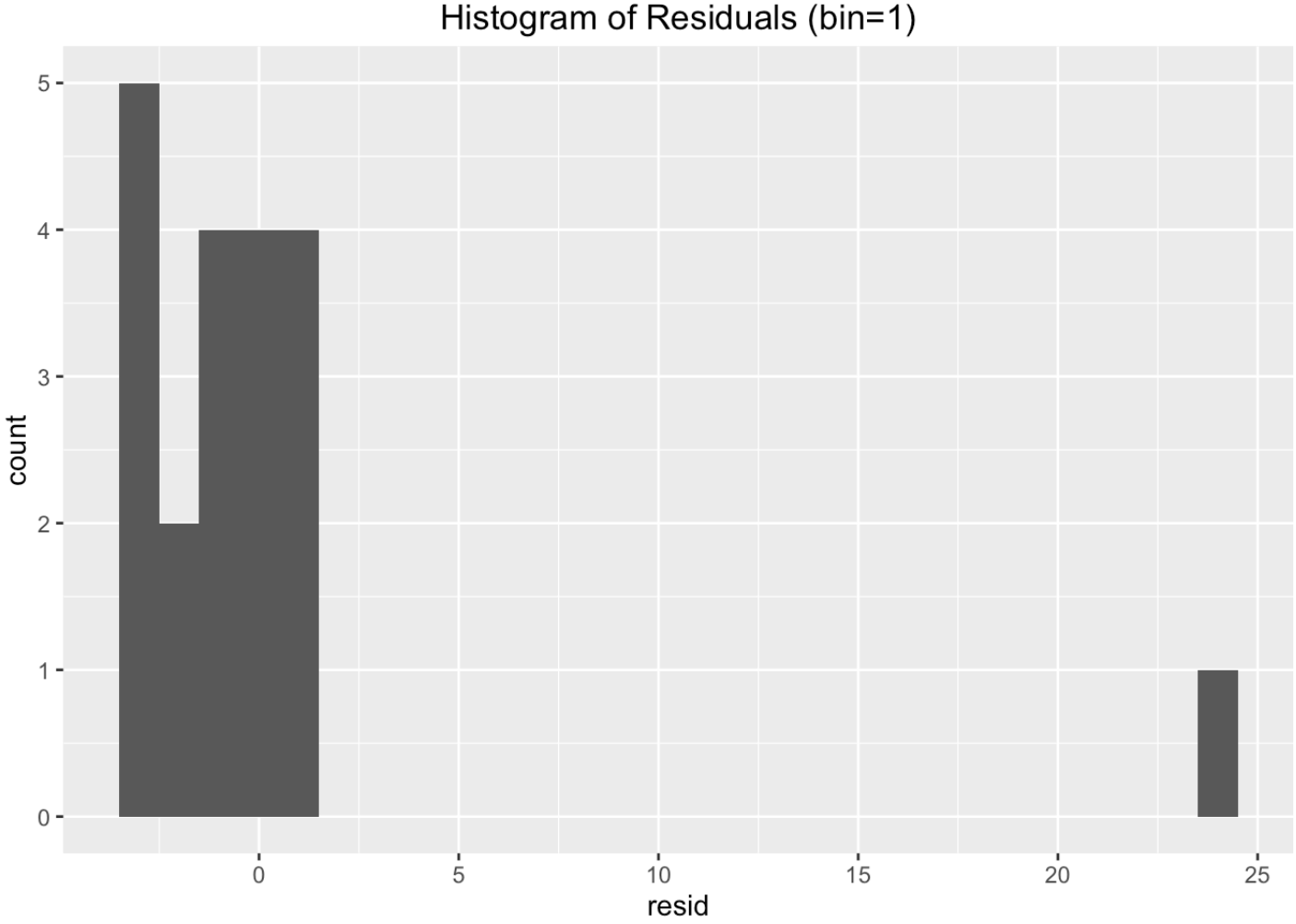


Estimate the coefficients a and b in the simple regression model with sales as dependent variable and advertising as explanatory factor. Also compute the standard error and t-value of b. Is b significantly different from 0?

Variable	Value
Coefficients	a = 26.5211, b = -0.0211
Std error	b = 0.2294
t-value	-0.09, not statistically significant from 0

Compute the residuals and draw a histogram of these residuals. What conclusion do you draw from this histogram?

```
residmodel = data.frame("resid" = resid(regmodel))
p = ggplot(residmodel, aes(resid)) +
  geom_histogram(binwidth = 1) +
  ggtitle("Histogram of Residuals (bin=1)")
p
```




```
summary(residmodel)
```

```
##      resid
##  Min.   :-3.41579
##  1st Qu.:-2.23158
##  Median :-1.32105
##  Mean   : 0.00000
##  3rd Qu.: 0.03947
##  Max.   :23.73158
```

Linear regression is not a good fit…residuals are not normally distributed due to the outlier data point.

Apparently, the regression result of part (b) is not satisfactory. Once you realize that the large residual corresponds to the week with opening hours during the evening, how would you proceed to get a more satisfactory regression model?

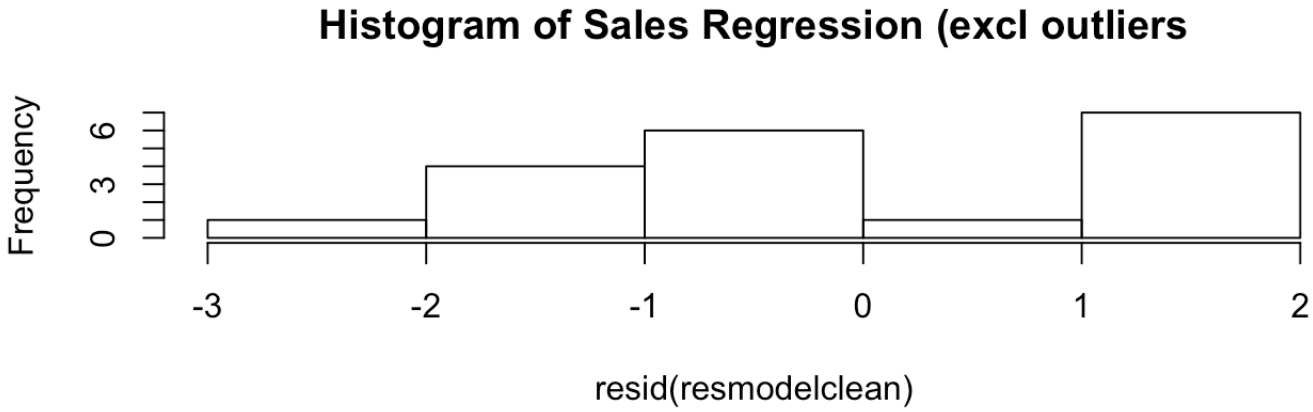
> You could remove the outlier to get cleaner data. This is acceptable because the data represents a point that was not operating under the same model as the other data points (open in the evenings)

Delete this special week from the sample and use the remaining 19 weeks to estimate the coefficients a and b in the simple regression model with sales as dependent variable and advertising as explanatory factor. Also compute the standard error and t-value of b. Is b significantly different from 0?

```
##
## Call:
## lm(formula = tsalesdatclean$Sales ~ tsalesdatclean$Observ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4733 -1.0853 -0.2405  1.2983  1.9147
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           25.86132    0.66726  38.758   <2e-16 ***
## tsalesdatclean$Observ -0.07760    0.05571  -1.393    0.182
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.434 on 17 degrees of freedom
## Multiple R-squared:  0.1025, Adjusted R-squared:  0.04966
## F-statistic: 1.941 on 1 and 17 DF,  p-value: 0.1816
```

Variable	Value
Coefficients	a = 25.86, b = -0.0776
Std error	b = 0.0557
t-value	-1.39, not statistically significant from 0

Discuss the differences between your findings in parts (b) and (e). Describe in words what you have learned from these results.



Removing the outlier data point significantly reduced the standard error, but the model is still not very descriptive. This is probably due to the fact that there are other external factors that are influencing sales