HW3

Nathan Krieger

2026-02-06

```r
library(ISLR2)
#head(Carseats)
```

Problem 1

(a)

(i)

```r
head(Carseats)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1  9.50       138     73          11        276   120       Bad  42        17
## 2 11.22       111     48          16        260    83      Good  65        10
## 3 10.06       113     35          10        269    80    Medium  59        12
## 4  7.40       117    100           4        466    97    Medium  55        14
## 5  4.15       141     64           3        340   128       Bad  38        13
## 6 10.81       124    113          13        501    72       Bad  78        16
##   Urban  US
## 1   Yes Yes
## 2   Yes Yes
## 3   Yes Yes
## 4   Yes Yes
## 5   Yes  No
## 6    No Yes
```

```r
m1 <- lm(Sales ~ Sales + CompPrice + Income +Advertising + Population + Price + Age + Education + Urban
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared on
## the right-hand side and was dropped
```

```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 1 in
## model.matrix: no columns are assigned
```

```r
summary(m1)$coefficients
```

```
##                  Estimate  Std. Error     t value     Pr(>|t|)
## (Intercept)  7.8243876204 1.129879216   6.9249770 1.805626e-11
## CompPrice    0.0942544671 0.007863742  11.9859556 2.131214e-28
## Income       0.0130501085 0.003490846   3.7383797 2.129338e-04
## Advertising  0.1369398910 0.021039352   6.5087503 2.332747e-10
## Population  -0.0002007219 0.000701023  -0.2863272 7.747796e-01
## Price       -0.0924395081 0.005062126 -18.2610052 2.781560e-54
## Age         -0.0447919380 0.006022554  -7.4373656 6.593817e-13
## Education   -0.0423034309 0.037373764  -1.1319018 2.583712e-01
## UrbanYes    -0.1559035988 0.213351937  -0.7307344 4.653802e-01
## USYes       -0.1062926485 0.283219156  -0.3753018 7.076401e-01
```

```r
#summary_table <- summary(m1)$coefficients
#print(summary_table)
```

(ii)

Null hypothesis (H0): The CompPrice of the car seat has zero effect on the sale of the car seats

Alternative hypothesis (H1): The CompPrice of the car seat has an effect on the sale of the car seats

Test statistic:

```r
comp_price_t_val <- summary(m1)$coefficients["CompPrice", "t value"]

print(comp_price_t_val)
```

```
## [1] 11.98596
```

```r
comp_price_p_val <- summary(m1)$coefficients["CompPrice", "Pr(>|t|)"]

print(comp_price_p_val)
```

```
## [1] 2.131214e-28
```

Since the p-value is less than $\alpha = 0.05$ (2.131214e-28), we reject the null hypothesis and conclude that CompPrice has a statistically significant relationship with car seat Sales.

TODO: NULL DISTRIBUTION

(b)

I needed to assume the ϵ (errors) are normally distributed.

(c)

```r
summary(m1)$sigma^2
```

```
## [1] 3.733413
```

$\hat{\sigma}^2 = 3.732624$

This is the estimated variance of the error term. It represents the average squared deviation of the actual sales values from the predicted regression line.

(d)

```r
summary(m1)$coefficients["Advertising", ]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 1.369399e-01 2.103935e-02 6.508750e+00 2.332747e-10
```

(e)

(f)

(g)

(h)

(i)

(j)

(k)

Problem 2

(a)

$m * \alpha$

(b)

```r
set.seed(123)
n = 1000  # Number of observations
p_test = c(200, 400, 500, 600, 800)
alpha = 0.05

false_positives <- c()

for (p in p_test) {
  x <- matrix(rnorm(n * p), n, p)
  y <- rnorm(n)
```

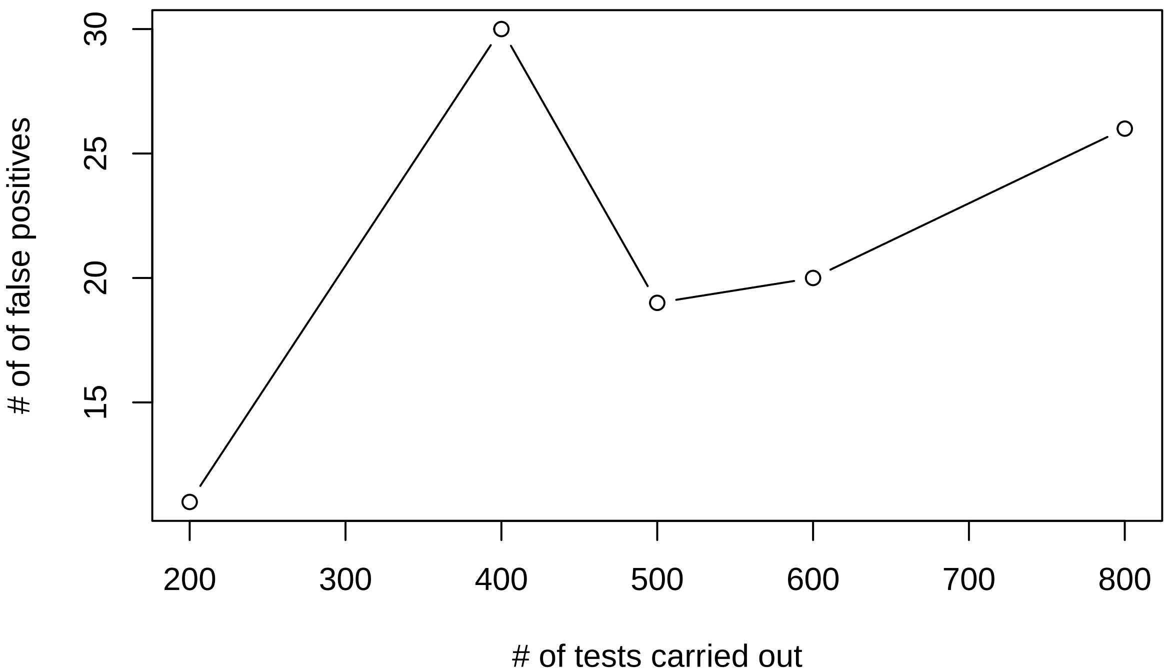
```r
  data <- as.data.frame(x)

  model <- lm(y ~ ., data = data)

  #Step 6: Extract p-values for all predictors
  p_values <- summary(model)$coefficients[-1,4]  # use -1 to Remove intercept

  #Step 7: Count number of significant predictors at 0.05 level
  significant_count = sum(p_values < alpha)
  # or
  false_positives <- c(false_positives, significant_count)
}

plot(p_test, false_positives, type="b",
     xlab = "# of tests carried out",
     ylab = "# of of false positives"
)
```



Problem 3

(a)

$\hat{\beta}_0 = 2$ $\hat{\beta}_1 = 3$ $\hat{\beta}_2 = 5$

(b)

```r
X1 = seq(0,10,length.out =100) #generates 100 equally spaced values from 0 to 10.
X2 = runif(100) #generates 100 uniform values.
```

(c)

(d)

(e)

Problem 4

(a)

In plain language, setting the significance level to a = 0.05 means that before we look at the data, we are deciding that we're willing to accept a 5% chance of making a false alarm.

(b)

I would disagree with that claim. This is because not significant is not the same as no effect.

(c)