ProblemSet2

Aaron Graybill

2021-03-02

Contents

Problem 2.1

```r
df <-
  read.csv(here('Data',"playbill.csv"))
movie_lm <-
  lm(CurrentWeek~LastWeek,data=df)
summary(movie_lm)
```

```
##
## Call:
## lm(formula = CurrentWeek ~ LastWeek, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -36926  -7525  -2581   7782  35443
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.805e+03  9.929e+03   0.685    0.503
## LastWeek    9.821e-01  1.443e-02  68.071   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18010 on 16 degrees of freedom
## Multiple R-squared:  0.9966, Adjusted R-squared:  0.9963
## F-statistic:  4634 on 1 and 16 DF,  p-value: < 2.2e-16
```

2.1.a

We can find the 95% confidence interval on the $\hat{\beta}_1$ term with the following code:

```r
confint_2.1 <-
  confint(movie_lm)
```

The confidence interval around $\hat{\beta}_1$ is $(0.9514971, 1.0126658)$ which actually contains 1, so we do not have significant evidence that the true $\beta_1 \neq 1$. We can compute this test a different way. We can run a hypothesis test where $H_0\colon \beta_1 = 1$ and $H_a\colon \beta_1 \neq 1$. We showed in class that:

$$\frac{\hat{\beta}_1 - 1}{se(\hat{\beta}_1)} \sim t_{n-2}$$

We have all of these values from our regression output, so we can compute the p-value that the null hypothesis is true in the following way:

```
test_stat=(movie_lm[["coefficients"]][["LastWeek"]]-1)/coef(summary(movie_lm))[, "Std. Error"]["LastWeel

critical_region <-
  qt(c(.025,.975),movie_lm$df.residual)

p_val <- (pt(test_stat,movie_lm$df.residual)*2)
```

The 95% critical region, as computed is $(-\infty, -2.12) \cup (2.12, \infty)$, and our test statistic is equal to -1.2419935 is not in that region, so we have insufficient evidence to reject the null hypothesis that $\beta_1 = 1$ at the 95% level. In fact, the p-value from this test is 0.2321368 which is well above the required .05 at the 95% level.

2.1.b

As proven in class:

$$\frac{\hat{\beta}_0 - \beta_0^0}{se(\hat{\beta}_0)} \sim t_{n-2}$$

In this case our H_0 is $\beta_0 = 1000$ and H_a is $\beta_0 \neq 1000$. Implementing similar code to above gives:

```
test_stat=(movie_lm[["coefficients"]][["(Intercept)"]]-1000)/coef(summary(movie_lm))[, "Std. Error"]["(I

critical_region <-
  qt(c(.025,.975),movie_lm$df.residual)

p_val <- (pt(-abs(test_stat),movie_lm$df.residual)*2)
```

I have to do some business with `-abs(test_stat)` to ensure that when I compute the cumulative density it's in the left tail so the p-value is just two times the computed density. Anyway, computing the p-value gives: 0.5669558 which is quite high and provides very little evidence that the true $\beta_0 \neq 1000$. In fact since $p > .5$, there is more evidence that H_0 is true than the alternative.

2.1.c

We showed in class that the prediction interval is given by:

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\frac{\alpha}{2},n-2} S \sqrt{\frac{1}{n} + \frac{x^* - \overline{x}}{SXX}}$$

But thankfully `R`will take care of that for us with the following code:

```
prediction_points <-
  data.frame(LastWeek=400000)
prediction <-
  predict(movie_lm,prediction_points,interval = 'prediction',level=.95,se.fit=T)
prediction
```

```
## $fit
##        fit      lwr      upr
## 1 399637.5 359832.8 439442.2
##
```

```
## $se.fit
## [1] 5318.889
##
## $df
## [1] 16
##
## $residual.scale
## [1] 18007.56
```

Summarizing those results, the point estimate for `CurrentWeek` is \$399637.5 with a 95% prediction interval of: $(359832.8, 439442.2)$. So we are 95% percent certain the true value of y^* would lie in the aforementioned range.

2.1.d

This heuristic is more or less appropriate. The regression coefficient is $\hat{\beta}_1 = 0.9820815$ which says that sales next week will be approximately 98% what they were last week which is quite close to exactly what they were last week. That 2% difference might not be a problem for some, but the true estimate is not exactly one. In fact, we could not conclusively show that the β_1 was different from 1, the p-value on that test was .23, not significant evidence to the contrary.

Problem 2.2

```r
df <-
  read.delim(here('Data',"indicators.txt"),sep = '\t')
```

2.2.a

I create the linear model and the confidence in the following way:

```r
econ_lm <-
  lm(PriceChange~LoanPaymentsOverdue,data=df)
summary(econ_lm)
```

```
##
## Call:
## lm(formula = PriceChange ~ LoanPaymentsOverdue, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6541 -3.3419 -0.6944  2.5288  6.9163
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.5145     3.3240   1.358   0.1933
## LoanPaymentsOverdue  -2.2485     0.9033  -2.489   0.0242 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.954 on 16 degrees of freedom
## Multiple R-squared:  0.2792, Adjusted R-squared:  0.2341
## F-statistic: 6.196 on 1 and 16 DF,  p-value: 0.02419
```

```r
confint(econ_lm)
```

```
##                         2.5 %     97.5 %
```

```
## (Intercept)          -2.532112 11.5611000
## LoanPaymentsOverdue -4.163454 -0.3335853
```

The 95% confidence interval on the $\hat{\beta}_1$ does not contain any positive values so we can be confident that the true slope, β_1, also is not positive. We could also run a hypothesis test to the same effect. Interpreting this, we can be reasonably certain that an increase in overdue loan payments is associated with a decrease in prices.

2.2.b

Here we are not doing a prediction interval, we are doing a confidence interval on the expected value of Y given $x = 4$. We implement that in the following way:

```r
data <- data.frame(LoanPaymentsOverdue=4)
predict(econ_lm,data,interval = 'confidence',se.fit=T,level = .95)
```

```
## $fit
##         fit       lwr       upr
## 1 -4.479585 -6.648849 -2.310322
##
## $se.fit
## [1] 1.023283
##
## $df
## [1] 16
##
## $residual.scale
## [1] 3.953998
```

The expected value of change in prices is -4.48% when $x = 4$ which is the percentage of loans overdue. The confidence interval does not include zero, so we can be reasonably sure that the expected value of price change is not zero.

Problem 2.4

2.4.a

We wish to minimize the square residuals and solve for the $\hat{\beta}$ that does so.

The sum of the square residuals are $\sum_i^n (y_i - \hat{y}_i)^2$. And our model is of the form: $\hat{y}_i = \hat{\beta} x_i$ so we have:

$$\arg\min_{\hat{\beta}} \left\{ \sum_i^n \left(y_i - \hat{\beta} x_i \right)^2 \right\}$$

The first order condition would then be:

$$\sum_i^n 2 \left(y_i - \hat{\beta} x_i \right) x_i = 0$$

Doing some algebraic manipulations gives:

$$
\begin{aligned}
\sum_i^n 2\left(y_i - \hat{\beta}x_i\right)x_i &= 0 \\
2\sum_i^n x_iy_i - \hat{\beta}x_i^2 &= 0 \\
\sum_i^n x_iy_i - \hat{\beta}x_i^2 &= 0 \\
\sum_i^n x_iy_i - \sum_i^n \hat{\beta}x_i^2 &= 0 \\
\sum_i^n x_iy_i - \hat{\beta}\sum_i^n x_i^2 &= 0 \\
\sum_i^n x_iy_i &= \hat{\beta}\sum_i^n x_i^2 \\
\frac{\sum_i^n x_iy_i}{\sum_i^n x_i^2} &= \hat{\beta}
\end{aligned}
$$

2.4.b

i.

$$
\begin{aligned}
E[\hat{\beta}|X = x_i] &= \\
&= E\left[\frac{\sum_i^n x_iy_i}{\sum_i^n x_i^2}|X = x_i\right] \\
&= \frac{E\left[\sum_i^n x_iy_i|X = x_i\right]}{E\left[\sum_i^n x_i^2|X = x_i\right]} \\
&= \frac{\sum_i^n x_iE\left[y_i|X = x_i\right]}{\sum_i^n x_i^2}\text{(conditioning on } x\text{)} \\
&= \frac{\sum_i^n x_i\beta x_i}{\sum_i^n x_i^2}\text{(by assumption)} \\
&= \beta\frac{\sum_i^n x_i^2}{\sum_i^n x_i^2} \\
&= \beta
\end{aligned}
$$

ii.

$$
\begin{aligned}
Var[\hat{\beta}|X = x_i] = & \\
&= Var\left[\frac{\sum_i^n x_i y_i}{\sum_i^n x_i^2}|X = x_i\right] \\
&= \frac{1}{\left(\sum_i^n x_i^2\right)^2} Var(\sum_i^n x_i y_i |X = x_i) \quad \text{(conditioning on } x\text{)} \\
&= \frac{1}{\left(\sum_i^n x_i^2\right)^2} \sum_i^n x_i^2 Var(y_i|X = x_i) \quad \mathbf{1.} \\
&= \frac{1}{\left(\sum_i^n x_i^2\right)^2} \sum_i^n x_i^2 Var(\beta x_i + e_i|X = x_i) \quad \text{(modelling assumption)} \\
&= \frac{1}{\left(\sum_i^n x_i^2\right)^2} \sum_i^n x_i^2 \sigma^2 \quad \mathbf{2.} \\
&= \frac{\sigma^2 \sum_i^n x_i^2}{\left(\sum_i^n x_i^2\right)^2} \\
&= \frac{\sigma^2}{\sum_i^n x_i^2}
\end{aligned}
$$

The **1**. step uses the conditioning on $X = x_i$ and the fact that the Y_is are independent. Step **2**. uses the conditioning on x coupled with the fact that we then only have a location shift so the variance is unchanged and finally the modeling assumption that $Var(e_i|X = x_i) = \sigma^2$.

iii. We have already shown that the mean and variance of $\hat{\beta}$ are as desired, now just to prove normality. As shown $\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$. Conditioning on X means the denominator is a constant. Furthermore, Conditioning on X means that the numerator is the weighted sum of a series of normal distribution (because $y|X \sim N$). Since all of the y_is are uncorrelated (and independent) by assumption, this weighted sum of normals must remain a normal distribution. Therefore, we have a normal distribution divided by a constant which itself must be a normal. Therefore, we have proven that $\hat{\beta}$ is distributed normally and since we already know its two parameters, we can fully characterize the distribution of $\hat{\beta}$ as $\hat{\beta} \sim N\left(\beta, \frac{\sigma^2}{\sum_i^n x_i^2}\right)$. ■