Predicting Final Exam Scores

Eduardo Martinez

Type : Homework Problem
Date Completed : 9/12/2021

Course : Applied Statistics/Regression (MATH-564)
Institution : Illinois Institute of Technology

```r
# Packages Used:
library(knitr)
library(kableExtra, warn.conflicts = F)
library(tidyverse, warn.conflicts = F)
```

The Data

```r
table3.10 <- read_tsv("Table3.10.txt", show_col_types = F)[c(2,3,1)]
Table3.10 <- cbind(Index = 1:22, table3.10)
colnames(Table3.10) <- c("Index", "$P_1$", "$P_2$", "$F$")
```

```r
kbl(cbind(Table3.10[1:11,], Table3.10[12:22,]), booktabs = T, escape = F,
    align = "c", linesep = "", valign = "c",
    caption = "$\\textbf{Table 3.10} - \\text{Examination Data}$") %>%
  kable_classic() %>%
  kable_styling(latex_options = c("condensed", "striped", "HOLD_position"), font_size = 11) %>%
  column_spec(c(2:4, 6:7), width = "1.25cm") %>%
  column_spec(c(1,5), width = "1.5cm", border_left = T, border_right = F) %>%
  column_spec(8, width = "1.25cm", border_right = T)
```

Table 3.10 – Examination Data

Index	P_1	P_2	F	Index	P_1	P_2	F
1	78	73	68	12	79	75	75
2	74	76	75	13	89	84	81
3	82	79	85	14	93	97	91
4	90	96	94	15	87	77	80
5	87	90	86	16	91	96	94
6	90	92	90	17	86	94	94
7	83	95	86	18	91	92	97
8	72	69	68	19	81	82	79
9	68	67	55	20	80	83	84
10	69	70	69	21	70	66	65
11	91	89	91	22	79	81	83

Exercise 3.3

Table 3.10 shows the scores in the final examination F and the scores in two preliminary examinations P_1 and P_2 for 22 students in a statistics course

---

(a) Fit each of the following models to the data:

$$\begin{aligned}
\text{Model 1:} \quad & F = \beta_0 + \beta_1 P_1 \qquad\quad + \varepsilon \\
\text{Model 2:} \quad & F = \beta_0 \qquad\quad + \beta_2 P_2 + \varepsilon \\
\text{Model 3:} \quad & F = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + \varepsilon
\end{aligned}$$

```r
fit_P1 <- lm(`F` ~ P1, data = table3.10)
coefP1 <- round(coefficients(fit_P1), 3)

fit_P2 <- lm(`F` ~ P2, data = table3.10)
coefP2 <- round(coefficients(fit_P2), 3)

fit_P1P2 <- lm(`F` ~., data = table3.10)
coefP1P2 <- round(coefficients(fit_P1P2), 3)
```

$$\begin{aligned}
\text{Fitted Model 1:} \quad & \hat{F} = \text{-22.342} + 1.261\ P_1 \\
\text{Fitted Model 2:} \quad & \hat{F} = \text{-1.854} \qquad\qquad + 1.004\ P_2 \\
\text{Fitted Model 3:} \quad & \hat{F} = \text{-14.501} + 0.488\ P_1 + 0.672\ P_2
\end{aligned}$$

(b) Test whether $\beta_0 = 0$ in each of the three models.

I will use t-test hypothesis test for each model where $H_0 : \hat{\beta}_0 = 0$ and $H_A : \hat{\beta}_0 \neq 0$.

There are $n = 22$ rows in the dataset. Under the null, the critical t-value has $n - p$ degrees of freedom ($d.f.$), where p equals the number of coefficients in the alternative regression model. Equivalently, p equals the number of predictors in a regression model since the intercept term is removed under the null.

Thus, Model 1 and Model 2 both have 20 $d.f.$ and Model 3 has 19 $d.f.$

Using a significance level, $\alpha = 0.05$, then the critical t-values for a two-tailed test are the following:

$$t_{(\alpha/2,\ d.f.=20)} = \pm 2.086 \qquad \text{and} \qquad t_{(\alpha/2,\ d.f.=19)} = \pm 2.093$$

Next, the following equation is used to calculate the test statistic for $H_A :\ t^* = \dfrac{\hat{\beta}_0 - 0}{s.e.(\hat{\beta}_0)}$.

We reject H_0 in favor of H_A if $|t^*| > |t_{(\alpha/2,\ d.f.)}|$.

```r
# Saving Model Summaries
sumP1 <- summary(fit_P1)
sumP2 <- summary(fit_P2)
sumP1P2 <- summary(fit_P1P2)
# Obtaining Standard Errors
seP1B0 <- sumP1$coefficients[1,2]
seP2B0 <- sumP2$coefficients[1,2]
seP1P2B0 <- sumP1P2$coefficients[1,2]
```

$$\begin{aligned}
\text{Model 1:} \quad |t^*| &= \left|\frac{-22.342}{11.564}\right| = 1.932 \; < 2.086 = |t_{(0.025,\ 20)}| \\
\text{Model 2:} \quad |t^*| &= \left|\frac{-1.854}{7.562}\right| = 0.245 \; < 2.086 = |t_{(0.025,\ 20)}| \\
\text{Model 3:} \quad |t^*| &= \left|\frac{-14.501}{9.236}\right| = 1.570 \; < 2.093 = |t_{(0.025,\ 19)}|
\end{aligned}$$

Conclusion

In all three models $|t^*| < |t_{(\alpha/2,\ d.f.)}$. As a result, we fail to reject the null hypothesis for all models. There is insufficient evidence in favor of the alternative hypothesis.

(c) Which Predictor is Better? P_1 or P_2? (Quick Model Selection)

The regression summaries for Model 1 and Model 2 are provided in the tables below:

```r
coefTab1 <- cbind(IV = c("(Intercept)", "$P_1$"), as_tibble(signif(coef(sumP1), 5)))
coefTab1[,5] <- as.character(signif(coefTab1[,5], 2))
kbl(coefTab1, booktabs = T, align = "c", escape = F, digits = 2, valign = "c",
    col.names = c(" ", "$\\widehat{\\beta}_j$", "$s.e.$", "$t^{\\ast}$", "$\\textit{p-value}$")) %>
  kable_classic() %>%
  kable_styling(latex_options = c("striped", "HOLD_position")) %>%
  add_header_above(c("Model 1" = 5), bold = T, font_size = 12) %>%
  column_spec(1:5, width = "2cm")
```

Model 1

	$\widehat{\beta}_j$	$s.e.$	t^*	*p-value*
(Intercept)	-22.34	11.56	-1.93	0.068
P_1	1.26	0.14	9.01	1.8e-08

Model 2

	$\widehat{\beta}_j$	$s.e.$	t^*	*p-value*
(Intercept)	-1.85	7.56	-0.25	0.81
P_2	1.00	0.09	11.09	5.4e-10

Both predictors are statistically significant as their *p-values* are less than the desired level of significance, $\alpha = 0.05$.

A quick way too access which predictor is better is comparing the R^2 and Mean Squared Error (MSE) statistics of each model. R^2 measures a model's goodness-of-fit; MSE is statistic used to evaluate the prediction accuracy of a model.

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \qquad \text{and} \qquad \text{MSE} = \frac{\text{SSE}}{n} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where $\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the sum of squares in total, n is the number of rows in the data ($n = 22$), and b is the number of regression coefficients in a model.

For both Models 1 and 2, there is an intercept and one predictor so $b = 2$; Model 3 has one addition coefficient so $b = 3$

- $0 \leq R^2 \leq 1$ such that R^2 is optimized when it is maximized.
- $\text{MSE} \geq 0$ such that MSE is optimized when it is minimized.

```r
ModStats <- tibble(" " = c("Model 1", "Model 2"),
                   `$R^2$ `= c(sumP1$r.squared, sumP2$r.squared),
                   `$\\textit{MSE}$` = c(mean(sumP1$residuals^2), mean(sumP2$residuals^2)))
```

The values for both statistic are provided in the table below:

	R^2	*MSE*
Model 1	0.8023	23.47
Model 2	0.8600	16.61

Conclusion:

This quick model selection method indicates that R^2 is maximized and *MSE* is minimized by Model 2. Therefore, I would prefer to use the second preliminary exam, P_2, to predict final exam scores, F.

(d) Which of the three models with intercepts would you use to predict the final examination scores for a student who scored 78 and 85 on the first and second preliminary examinations, respectively? (Quick Model Selection) What is your prediction in this case?

R^2 becomes larger as more predictors are used to fit a model. This means R^2 does not account the bias of larger models.

In contrast, adjusted R^2 denoted R^2_{adj} accounts for bias by punishing models as they add predictors:

$$R^2_{adj} = 1 - \frac{\text{SSE}/(n-b)}{\text{SST}/(n-1)} = 1 - \frac{\text{SSE}\cdot(n-1)}{\text{SST}\cdot(n-b)}.$$

The following properties of R^2 still hold: $0 \leq R^2_{adj} \leq 1$ such that R^2_{adj} is optimized when it is maximized. However, R^2_{adj} decreases as predictors are added when all other values are fixed.

Because Model 3 has an additional coefficient, R^2_{adj} should be used to compare it to the smaller models instead of the unadjusted R^2.

```
ModStats2 <- tibble(" " = c("Model 1", "Model 2", "Model 3"),
              `$R^2_{adj}$ `= c(sumP1$adj.r.squared, sumP2$adj.r.squared, sumP1P2$adj.r.squared),
              `$\\textit{MSE}$` = c(mean(sumP1$residuals^2), mean(sumP2$residuals^2), mean(sumP1P2$
```

The values of R^2_{adj} and MSE for each model are displayed in the table below:

	R^2_{adj}	MSE
Model 1	0.7924	23.47
Model 2	0.8530	16.61
Model 3	0.8744	13.49

Conclusion:

R^2_{adj} and MSE were optimized by *Model 3*. Recall, the estimated coefficients for *Model 3* derived in Part (a):

- $\hat{F}$ = -14.501 + 0.488 P_1 + 0.672 P_2

Accordingly, if a student had preliminary examination scores $P_1 = 78$ and $P_2 = 85$, *Model 3* predicts this student will have a final examination score of $\boxed{\hat{F} = 80.713}$.