

KEY

Stat 328 Final Exam (Regression)

Summer 2002

Professor Vardeman

This exam concerns the analysis of 1990 salary data for $n = 30$ offensive backs in the NFL. (This is a part of the larger data set that serves as the basis of your Lab #6.) Attached to this exam are a number of JMP reports for these data. Use them in answering the questions on this exam.

As on Lab #6, the variables available for modeling were:

<i>salary</i>	1990 season salary
<i>draft</i>	round in the player draft when the player was selected
<i>yrs_exp</i>	years of NFL experience for the player
<i>played</i>	the number of regular season games played in 1989
<i>started</i>	the number of regular season games started in 1989
<i>citypop</i>	the population of the city in which the player's team is located

Vardeman used these variables and created several more:

<i>log10salary</i>	the base 10 logarithm of <i>salary</i>
<i>1/draft</i>	the reciprocal of <i>draft</i>
<i>percentstarted</i>	the ratio <i>started/played</i>

To begin, consider the problem of modeling a salary variable in terms of only a draft position variable.

- a) Consider the plots of *salary* vs *draft* and *log10salary* vs *draft*. What about these suggests that (as far as using standard statistical methodology is concerned) *log10salary* is probably a better "y" than *salary*?

The "constant σ " SLR model assumption looks very dubious for *salary*. It appears to be less problematic for *log10salary*.

Ultimately, Vardeman decided to use *1/draft* instead of *draft* in a SLR regression analysis for *log10salary*. So until further notice, consider a SLR analysis using the model

$$\log_{10} \text{salary} = \beta_0 + \beta_1 (1/\text{draft}) + \varepsilon$$

- b) What fraction of raw variability in *log10salary* is accounted for using *1/draft* as a predictor variable?

$$R^2 = .266$$

- c) Give and interpret a *p*-value for testing $H_0 : \beta_1 = 0$. Say exactly where you found this on the printout.

p-value: .0035

where: in the ANOVA table, page 6

interpretation:

This is pretty small. We have strong evidence that mean *log10salary* changes with *1/draft*. There is statistically detectable linear relationship between these variables.

- d) Notice that if *draft* is large, $1/\text{draft}$ is near 0. So β_0 might be interpreted as a mean $\log_{10}\text{salary}$ for a high draft number (or perhaps even undrafted) offensive back. Give 95% confidence limits for this.

Here a CI for β_0 is interpretable. Use

$$b_0 \pm t s_{b_0}$$

$$5.36 \pm 2.048(.0788)$$

$$5.36 \pm .16$$

upper 2.5% of t_{28} distn

- e) What does the SLR model on the previous page give as the difference in mean $\log_{10}\text{salary}$ values for 1st and 2nd round draft picks? (Note that these are the cases $1/\text{draft} = 1$ and $1/\text{draft} = .5$.) Give 95% confidence limits for this difference in means.

This is $(\beta_0 + \beta_1(1)) - (\beta_0 + \beta_1(\frac{1}{2})) = \frac{1}{2}\beta_1$. Limits for β_1 are

$$b_1 \pm t s_{b_1} \quad \text{i.e.} \quad .46 \pm 2.048(.144)$$

$$\text{i.e.} \quad .46 \pm .29$$

So limits for the difference are $.23 \pm .15$

- f) A particular offensive back not included in this data set is a former first round draft pick and was offered a \$315,000 contract for 1990. (The base 10 logarithm of 315,000 is about 5.5.) On the basis of draft position alone, did this person have a good case that the offer was too low? Explain carefully.

No. While this offer is below the fitted mean for a first round back (and outside confidence limits for the mean) it is not outside prediction limits for $\log_{10}\text{salary}$ at $1/\text{draft} = 1$.

Now consider MLR analyses of $\log_{10}\text{salary}$. Notice that printouts are available for two different multiple linear regressions. The first is a regression on *draft*, *years_exp*, *played*, *started*, *citypop*, $1/\text{draft}$, and *percentstarted*. The second is a regression on only *years_exp*, $1/\text{draft}$, and *percentstarted*.

- g) Give 95% confidence limits for the standard deviation of $\log_{10}\text{salary}$ when all of *draft*, *years_exp*, *played*, *started*, *citypop*, $1/\text{draft}$, and *percentstarted* are held fixed.

These are confidence limits for σ . Use

$$\left(s_e \sqrt{\frac{n-k-1}{U}}, s_e \sqrt{\frac{n-k-1}{L}} \right)$$

$$\left(.183 \sqrt{\frac{22}{36.781}}, .183 \sqrt{\frac{22}{10.982}} \right)$$

$$(.14, .26)$$

h) What on the MLR printouts suggests that it may be feasible to model $\log_{10}\text{salary}$ using fewer than 7 predictors?

On page 7 many of the tests for individual coefficients have big p-values. Comparing pages 7 and 8, the decrease in R^2 isn't "huge", the increase in s_e is small, and PRESS actually decreases when 4 predictors are dropped from the full model. The Residual by Predicted plot on page 8 doesn't look any "worse" than the one on page 7.

i) There is a decrease in R^2 if one moves from the 7 variable regression to the 3 variable regression. Compute an appropriate F value, degrees of freedom and approximate p -value to attach to the decrease.

$$F = \frac{(SSR_{\text{full}} - SSR_{\text{red}}) / 4}{MSE_{\text{full}}} = \frac{(2.066357 - 1.909011) / 4}{.033347} = 1.18$$

$F: 1.18$

$df: 4, 22$

$p\text{-value: big}$

Henceforth consider the 3 variable regression. Besides the raw data, the JMP data table at the end of the printout has summaries of that fit. Notice that although $n = 30$ cases were used in the fitting, the table includes some values for an additional (31st) case.

j) According to this model, what increase in mean $\log_{10}\text{salary}$ accompanies a 1 year increase in NFL experience, if draft position and percentage of games started are held fixed? Give 95% confidence limits.

This is $\beta_{\text{years-exp}}$. Use

$$\begin{aligned} & b_{\text{years-exp}} \pm t S_{b_{\text{years-exp}}} \\ & .0512 \pm (2.056) .012582 \\ & .0512 \pm .0259 \end{aligned}$$

k) Dropping which of the 3 predictors would cause the biggest decrease in R^2 ? How do you know?

variable: years-exp

reasoning: It has the largest F (and t) value on page 8. These are for testing individual β 's equal to 0 in the 3 variable model and big F indicates big reduction in SSR and thus big reduction in R^2 .

- l) Player 30 has a large "hat" value. What about his values of *years_exp*, *1/draft*, and *percentstarted* makes this qualitatively plausible/expected?

Player 30 has the largest *years_exp* and smallest *percentstarted* in the data set. The *1/draft* value for this case, while not absolutely the smallest in the data set, is also fairly extreme. This case is "on the edge" of the cloud of z vectors.

- m) Considering both "x" and "y" variables, which player among the 30 in the data set was the "most influential" in terms of fitting the 3 variable model? Explain.

Case 1 has the largest Cook's D, combining a fairly big "hat" with a big deleted residual. One could expect that redoing the fitting without this case might substantially change the looks of the analysis.

- n) Make 95% prediction limits for the \log_{10} salary of player 31. (Use the 3 variable model!)

$$\text{Use } \hat{y} \pm t \sqrt{s_e^2 + s_m^2}$$

$$5.46866 \pm 2.056 \sqrt{(.185116)^2 + (.07184125)^2}$$

$$5.46866 \pm .4083$$

$$(5.0604, 5.8769)$$

- o) Player 31's actual salary was \$75,000. Does your answer to n) provide solid statistical evidence that his salary (for unknown reasons) was atypical? Explain.

$$\frac{5.0604}{10} = 115,000$$

The lower prediction limit for the salary for player 31 is substantially above the actual salary. The salary was (for whatever reason) clearly atypical.