

Stat 328 Lab #6 Summer 2002

On the Stat 328 "Data" web page, you will find a fairly large data set titled "Frees NFL Data." There is information on 137 veteran 1990 NFL football players taken from *Data Analysis Using Regression Models* by Edward Frees. Download those from the web page and enter them into JMP. (Right click on the link, download to your computer and open in JMP by choosing "Text Import Files" type of file and choosing the "attempt to discern format" option.) Recorded are values for:

"salary"	1990 season salary (\$)
"position"	1 = offensive back, 2 = defensive back, 3 = lineman, 4 = kicker/punter
"ob"	an indicator variable (1 = offensive back, 0 = other) (using this variable is an alternative to using the qualitative variable <i>position</i>)
"draft"	round of player draft in which the player was selected
"yrs_exp"	years of NFL experience for the player
"played"	the number of regular season games played in 1989
"started"	the number of regular season games started in 1989
"citypop"	the population of the city in which the player's team is located

Your fundamental task is to build an effective model for "salary" in terms of the other variables, (that might, for example, have been used in a 1990 salary arbitration case). You have at your disposal JMP, all you've learned about MLR, and your good business sense. You may:

- transform any variable(s) (for example, you might find the logarithm of salary to be more easily modeled than salary itself, the reciprocal of draft might be a better predictor than draft, etc.)
- make up new predictors from the ones above (for example you could allow for a different yearly increase in salary for offensive backs than other players by inventing a product variable (or what Dielman calls an interaction variable in Section 7.2) $ob \cdot yrs_exp$, or it might be useful to employ a ratio $started/played$ as a predictor)

Your search should consider issues such as parsimony/simplicity of the final model you settle on, the possibility that a few unusual cases dominate the fitting of the model and should really be eliminated from the fitting, goodness of fit criteria like R^2 , $PRESS$, MSE and C_p , sensible-looking residuals, etc. Be sure that if you use the qualitative variables above in your analysis, you handle them in a sensible manner (for example, you surely don't want to treat the codes for "position" as if they were values of a measured variable).

When you have settled on a model you'd be willing to defend in a court of law, compute and plot residuals against all of the predictors and against \hat{y} , and make a histogram of them (fitting a normal curve to the histogram using the "fit distribution" menu under the Analyze/Distribution procedure). Then make 95% prediction limits for all the cases you retain in your data set. (If you end up modeling a transform of y , compute the limits on the transformed scale and then "untransform" them to get limits on the original scale.) How do your predictions agree with the original y 's? Write a paragraph or two interpreting your model (to the extent that is possible) to the proverbial "layman." (What variables are the best predictors of salary, what cases if any did you delete from consideration and why, etc.)

There is (obviously) no "right" answer here. Do something sensible and document and defend it. On the 328 Summer 2001 page there are some comments that the grader made after looking at student solutions for this problem produced last summer. Take his comments into account as you prepare your solution.