Stat 506 Assignment 4

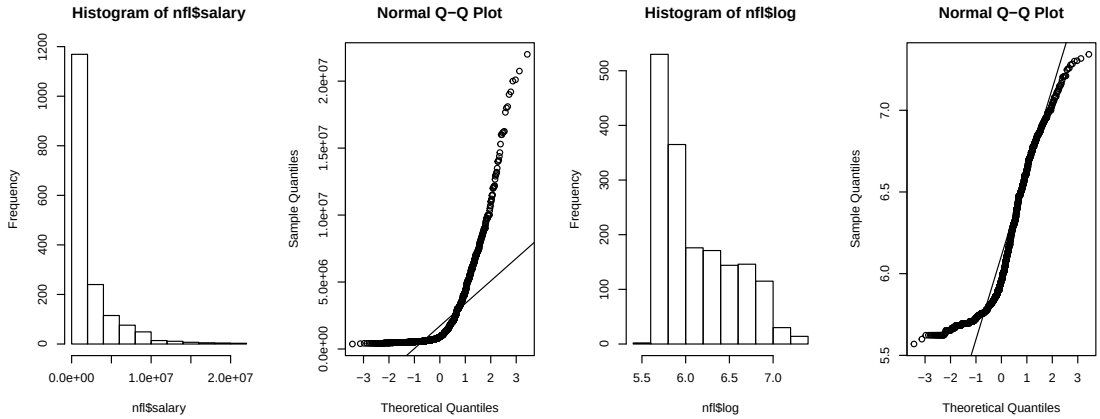
Leslie Gains-Germain

February 13, 2015

1. Im sure you know that salaries often exhibit a skewed distribution.

   (a) Investigate these data to see if a transformation is needed. If so, do it throughout this assignment. (Explain.)

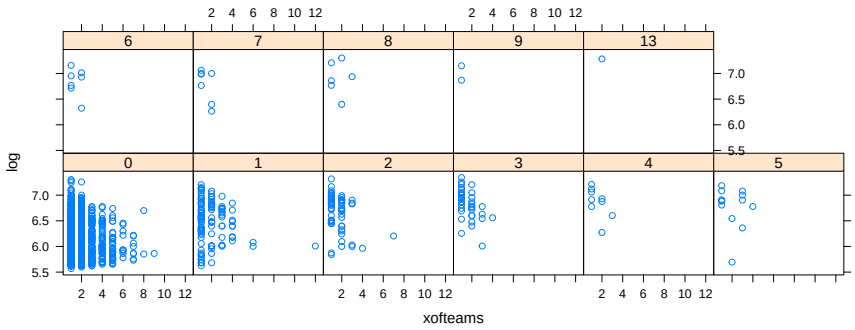
   The salaries look very skewed. I tried a log base 10 *transformation and the histogram of log salaries looks better. The distribution of log salaries is short tailed and closer to being symmetric than before the transformation. I chose to use a log base* 10 *transformation because it makes more sense to think of changes in salaries as factors of ten (rather than factors of e) after backtransforming.*



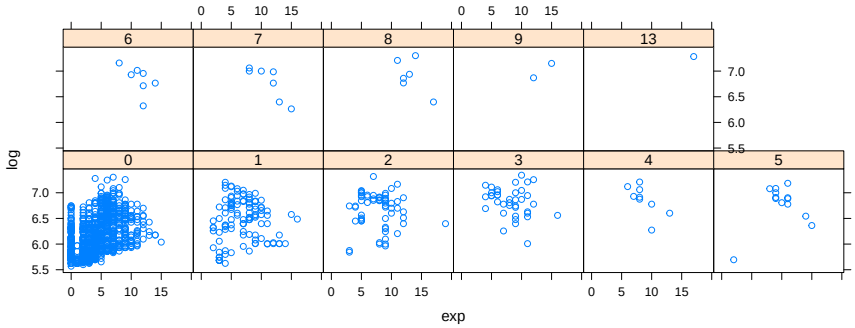
   (b) Decide which variables you will consider as predictors. Format them in a way that makes sense. (Which are factors?) Provide initial plots to see how they might relate to salary.

   I chose to include almost all of the variables as potential predictors. The only ones I didn't include were birth city, state, jersey number, high school, and high school city. I do not think the log salaries would depend greatly on these variables, and they would be factors with hundreds of levels (thus lots of computing power would be required to include them in the model).

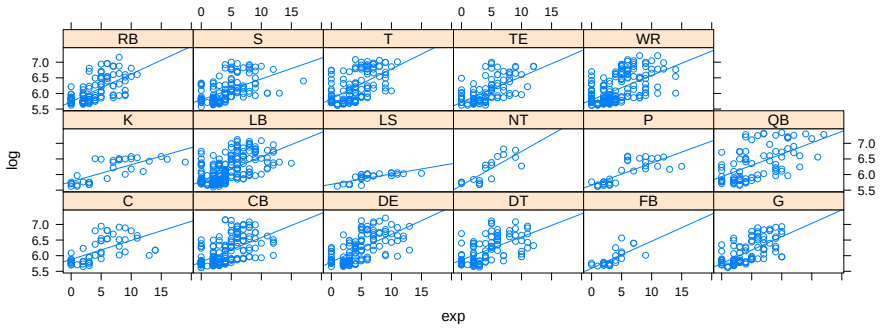
   In the plot below we see that a player's log salary generally decreases as the number of teams they have been on increases. This seems to be true across different numbers of probowl appearances.



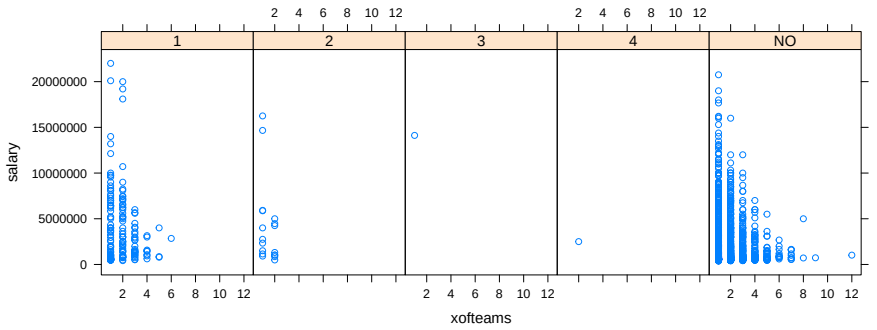
There does appear to be an interaction between probowl and experience. For those who have never won a probowl, more experienced people get higher salaries. For those people who have been in 1, 2, *or* 3 *probowls, there does not seem to be a clear relationship between log salary and experience. For those players who have been in four or more probowls, log salary is highest for players with the fewest years of experience. These must be the superstars! We can also see from this plot that people who have won more probowls have more years of experience, which make sense. The relationship between log salary and experience appears to be about the same across position, with the exception of the 'LS' position.*



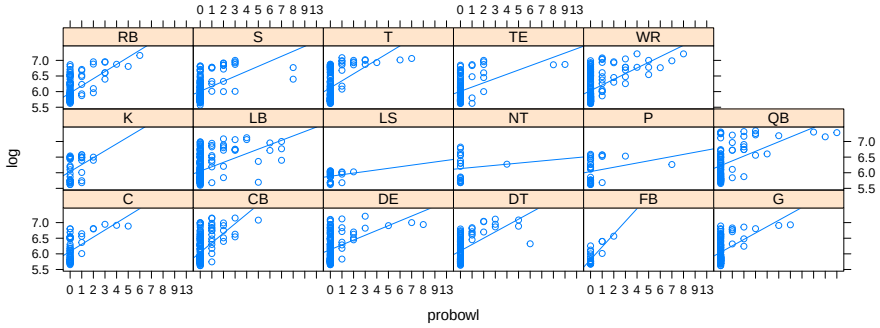
When we look at the relationship between log salary and experience across position, in every position we see a strong positive relationship between log salary and experience.



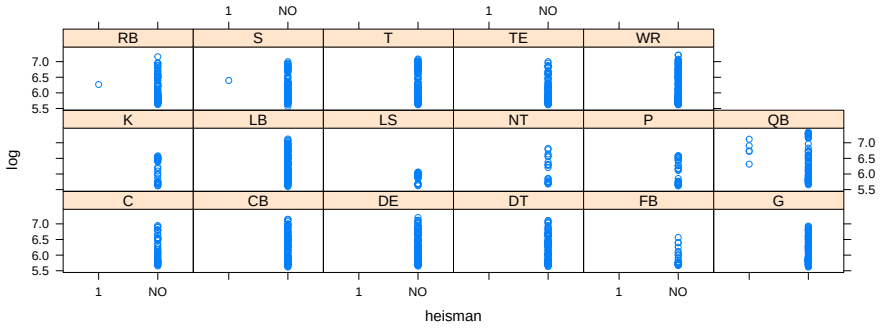
In this plot we can see that the players who have been champions more than once have only played on one or two teams.



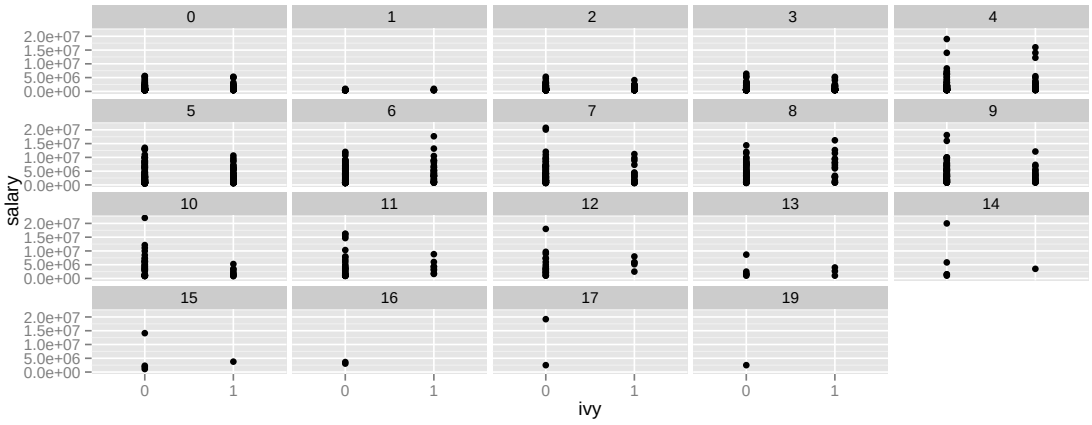
In the following plot we see that players who have been in more probowls generally have higher log salaries.



This plot shows that only seven people in the dataset have won the heisman, and five of them are quarterbacks. These seven have middle to high log salaries.



I didn't want to include college as a factor because there are 269 *different colleges in the dataset. To get around this issue, I made an indicator variable called 'ivy' that is coded as* 1 *if they went to a school that has a football program that is in the top* 25*. The following plot shows that a player's log salary does not depend much on whether or not they went to a top* 25 *rated college.*



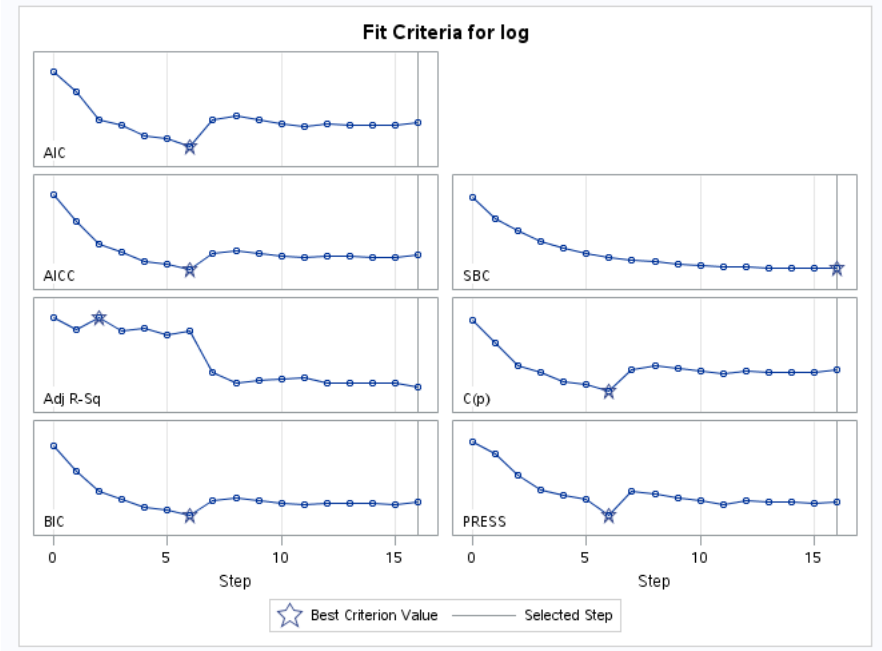
2. Demonstrate how the step function can be used in R to do forward, backward, or stepwise variable selection. Start with a full model to use backward and for one both directions fit. Start with just one variable to use forward and another both fit. Do the four results agree? What is the best AICc fit youve found?

   In R, I used the `step()` *function to choose a model, with the AICC criterion. The backward, forward, and stepwise selection procedures all chose a model with experience, probowl, number of teams, position, age, weight, height, heisman, probowl*number of teams, and experience*age as predictors. The AICC for this model is* -4122.26*. In the end, I chose this model but I eliminated the probowl*number of teams interaction. I did not include this interaction because there is not enough information in the dataset to estimate all probowl*number of teams interaction. When this interaction is removed, the AICC only increases by about 6 to* -4116.147*. I fit this linear model and the table of coefficients is shown below.*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.8091	0.3862	12.45	0.0000
posCB	0.1908	0.0775	2.46	0.0139
posDE	0.0522	0.0510	1.02	0.3056
posDT	0.0136	0.0510	0.27	0.7894
posFB	-0.0430	0.0786	-0.55	0.5842
posG	-0.0505	0.0507	-1.00	0.3190
posK	0.1402	0.0878	1.60	0.1104
posLB	0.0661	0.0565	1.17	0.2425
posLS	-0.2417	0.0787	-3.07	0.0022
posNT	-0.0295	0.0800	-0.37	0.7122
posP	0.0070	0.0856	0.08	0.9347
posQB	0.2263	0.0722	3.14	0.0017
posRB	0.1159	0.0696	1.67	0.0960
posS	0.0950	0.0729	1.30	0.1929
posT	0.0259	0.0500	0.52	0.6045
posTE	-0.0212	0.0607	-0.35	0.7274
posWR	0.1530	0.0746	2.05	0.0405
xofteams	-0.0600	0.0073	-8.23	0.0000
probowl1	0.2418	0.0306	7.89	0.0000
probowl2	0.3490	0.0407	8.58	0.0000
probowl3	0.4749	0.0506	9.39	0.0000
probowl4	0.4506	0.0999	4.51	0.0000
probowl5	0.3844	0.0843	4.56	0.0000
probowl6	0.3531	0.1147	3.08	0.0021
probowl7	0.3517	0.1149	3.06	0.0023
probowl8	0.4229	0.1276	3.31	0.0009
probowl9	0.4442	0.2129	2.09	0.0371
probowl13	0.7519	0.3038	2.48	0.0134
exp	0.2323	0.0165	14.10	0.0000
ht	0.0101	0.0048	2.09	0.0370
wt	0.0016	0.0006	2.63	0.0086
age	-0.0002	0.0076	-0.03	0.9738
heismanNO	-0.2116	0.1156	-1.83	0.0673
exp:age	-0.0050	0.0006	-9.00	0.0000

3. Read the data into SAS. Use the PROC GLMSELECT (2006?) to fit a variety of models. Refer to the example on modeling baseball salaries based on performance measures. Use forward, backward and stepwise. For at least one of these invoke stats = all and compare the different criterions favorite models.

   *I input the same candidate set into PROC GLMSELECT in SAS. I first used backward model selection (with the default SAS selection criterion, SBC). The model selected had number of teams, probowl, experience, weight, height*weight, and experience*age as predictors. We can see in the plot below that this model had the lowest SBC, but most of the other statistics (C_p, BIC, $AICC$, and $PRESS$) agreed on the model from step* 6 *(see steps below).*



The values of all criteria for the model, chosen by the SBC criteria with backward selection, are shown below.

The SAS System

The GLMSELECT Procedure

Backward Selection Summary

Step	Effect Removed	Number Effects In	Number Parms In	Model R-Square	Adjusted R-Square	AIC	AICC	BIC	CP	SBC	PRESS	ASE	F Value	Pr > F
0		23	219	0.5835	0.5220	-2286.3423	-2220.2825	-3914.3104	219.0000	-2791.2399	164.7753	0.0735		
1	hsstate	22	160	0.5626	0.5172	-2321.1874	-2287.1155	-3984.2772	175.2057	-3146.7062	162.3577	0.0772	1.26	0.0931
2	team	21	129	0.5583	0.5222*	-2366.7422	-2344.9368	-4038.3540	128.3177	-3360.7230	158.5143	0.0780	0.48	0.9929
3	exp*team	20	98	0.5441	0.5164	-2375.2920	-2362.8626	-4058.6346	116.4611	-3537.7348	155.6410	0.0805	1.62	0.0174
4	ht*pos	19	82	0.5407	0.5176	-2394.6641	-2385.9978	-4080.8616	96.5407	-3644.0550	154.6512	0.0811	0.75	0.7475
5	pos	18	66	0.5331	0.5145	-2398.7837	-2393.1763	-4088.5524	91.5318	-3735.1227	153.9671	0.0824	1.67	0.0456
6	wt*probowl	17	58	0.5324	0.5161	-2412.3625*	-2408.0269*	-4102.6927*	77.8968*	-3792.1756	150.8696*	0.0825	0.29	0.9691
7	wt*pos	16	42	0.5109	0.4988	-2368.2815	-2365.9868	-4063.1247	121.9618	-3835.0427	155.5276	0.0863	4.70	<.0001
8	xofteams*probowl	15	34	0.5042	0.4943	-2361.0588	-2359.5380	-4056.4587	129.8689	-3871.2941	154.9359	0.0875	2.85	0.0038
9	wt*race	14	30	0.5038	0.4951	-2367.5786	-2366.3842	-4062.7472	123.4038	-3899.5509	154.2324	0.0876	0.36	0.8352
10	race	13	26	0.5031	0.4957	-2373.5063	-2372.5982	-4068.4475	117.5551	-3927.2156	153.5853	0.0877	0.51	0.7290
11	champ	12	24	0.5031	0.4963	-2377.4635	-2376.6836	-4072.2567	113.5995	-3942.0413	152.8953	0.0877	0.02	0.9791
12	exp*champ	11	20	0.5000	0.4943	-2374.7452	-2374.1923	-4069.4841	116.7684	-3961.0601	153.5500	0.0883	2.65	0.0318
13	age	10	19	0.4997	0.4943	-2375.7525	-2375.2501	-4070.4260	115.8065	-3967.5016	153.4177	0.0883	0.98	0.3220
14	ivy	9	18	0.4994	0.4943	-2376.6166	-2376.1623	-4071.2255	114.9949	-3973.8000	153.3166	0.0884	1.12	0.2893
15	ht	8	17	0.4989	0.4941	-2377.0712	-2376.6626	-4071.6207	114.6132	-3979.6888	153.2535	0.0884	1.53	0.2163
16	heisman	7	16	0.4968	0.4923	-2372.1151	-2371.7497	-4066.6996	119.9155	-3980.1670*	153.3479	0.0888	6.90	0.0087

* Optimal Value of Criterion

Root MSE	0.29943
Dependent Mean	6.12059
R-Square	0.4968
Adj R-Sq	0.4923
AIC	-2372.11508
AICC	-2371.74971
BIC	-4066.69965
C(p)	119.91553
PRESS	153.34791
SBC	-3980.16697
ASE	0.08881

I also used forward selection and stepwise selection. These methods both selected dif-

ferent models. It seems like the main difference between the step procedure in SAS, compared to R, is that the models selected have interactions with effects that are not included as main effects in the model. I think this is the reason why the model selected by SAS varies a lot with the selection procedure.

4. Using a favorite model from number 3, fit in PROC MIXED with a random team intercept. Does this explain much more variance? Do effects look stronger as a result?

   Below on the left, I show the fixed effects from the model selected by SAS using forward selection. On the right, I show the estimates for the fixed effects when the model is fit with a random intercept for team. We can see that the estimates of the fixed effects are the same, but the standard errors are slightly smaller when a random intercept is allowed for team. As a result the p-values are slightly smaller and the effects do look stronger.

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	6.654519	0.309601	21.49
xofteams	1	-0.058580	0.007025	-8.34
probowl 0	1	-0.896781	0.307097	-2.92
probowl 1	1	-0.650273	0.307164	-2.12
probowl 2	1	-0.531502	0.307750	-1.73
probowl 3	1	-0.398999	0.309039	-1.29
probowl 4	1	-0.412258	0.321136	-1.28
probowl 5	1	-0.479801	0.315225	-1.52
probowl 6	1	-0.525690	0.323660	-1.62
probowl 7	1	-0.533989	0.323256	-1.65
probowl 8	1	-0.422814	0.325197	-1.30
probowl 9	1	-0.421368	0.368138	-1.14
probowl 13	0	0	.	.
exp	1	0.232908	0.016257	14.33
wt	1	-0.003830	0.001365	-2.81
ht*wt	1	0.000054063	0.000016090	3.36
exp*age	1	-0.005093	0.000493	-10.32

Solution for Fixed Effects

Effect	probowl	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		6.6545	0.3081	31	21.60	<.0001
xofteams		-0.05858	0.006992	1646	-8.38	<.0001
probowl	0	-0.8968	0.3056	1646	-2.93	0.0034
probowl	1	-0.6503	0.3057	1646	-2.13	0.0336
probowl	2	-0.5315	0.3063	1646	-1.74	0.0829
probowl	3	-0.3990	0.3076	1646	-1.30	0.1947
probowl	4	-0.4123	0.3196	1646	-1.29	0.1973
probowl	5	-0.4798	0.3137	1646	-1.53	0.1264
probowl	6	-0.5257	0.3221	1646	-1.63	0.1029
probowl	7	-0.5340	0.3217	1646	-1.66	0.0972
probowl	8	-0.4228	0.3237	1646	-1.31	0.1916
probowl	9	-0.4214	0.3664	1646	-1.15	0.2503
probowl	13	0	.	.	.	.
exp		0.2329	0.01618	1646	14.39	<.0001
wt		-0.00383	0.001358	1646	-2.82	0.0049
wt*ht		0.000054	0.000016	1646	3.38	0.0008
exp*age		-0.00509	0.000491	1646	-10.37	<.0001

5. Repeat number 4 in R using a favorite model from number 2. Explain how much difference you see between teams.

   The intercept varies very little across teams. The estimated σ^2_{team} *is* $1.80321 * 10^{-5}$*, and the estimated random effects for each team are shown below. We can see that the difference in intercept across teams is very small.*

	(Intercept)
49ers	0.0000000097
Bears	-0.0000000030
Bengals	0.0000000037
Bills	0.0000000146
Broncos	-0.0000000053
Browns	0.0000000028
Buccaneers	0.0000000029
Cardinals	-0.0000000067
Chargers	-0.0000000052
Chiefs	0.0000000121
Colts	0.0000000039
Cowboys	-0.0000000044
Dolphins	0.0000000098
Eagles	0.0000000164
Falcons	-0.0000000079
Giants	0.0000000042
Jaguars	0.0000000062
Jets	-0.0000000074
Lions	0.0000000051
Packers	0.0000000010
Panthers	-0.0000000101
Patriots	-0.0000000019
Raiders	-0.0000000150
Rams	-0.0000000001
Ravens	-0.0000000022
Redskins	-0.0000000065
Saints	-0.0000000056
Seahawks	-0.0000000072
Steelers	-0.0000000041
Texans	-0.0000000116
Titans	0.0000000051
Vikings	0.0000000070

Code Appendix

```
nfl <- read.csv("~/Documents/Stat506/Homework/HW4/nfl.csv")
nfl <- na.omit(nfl)
names(nfl) <- c("team", "x", "name" , "first", "last", "pos" ,"ht", "wt", "age", "bday", "exp", "college", "salary", "city","sta

dollarToNumber <- function(vector) {
  vector <- as.character(vector)
  vector <- gsub("(\\$|,)","", vector)
  result <- as.numeric(vector)
  return(result)
}

nfl$salary <- dollarToNumber(nfl$salary)
nfl <- subset(nfl, salary !=0)
```

```
par(mfrow=c(1,4))
hist(nfl$salary)
qqnorm(nfl$salary)
qqline(nfl$salary)
nfl$log <- log10(nfl$salary)
hist(nfl$log)
qqnorm(nfl$log)
```

```r
qqline(nfl$log)
```

```r
require(lattice)
nfl$pos <- factor(nfl$pos)
nfl$probowl <- factor(nfl$probowl)
xyplot(log~xofteams|probowl, data=nfl)
```

```r
xyplot(log~exp|probowl, data=nfl, type=c("p"))
```

```r
xyplot(log~exp|pos, data=nfl, type=c("p","r"))
```

```r
xyplot(salary~xofteams|champ, data=nfl, type=c("p"))
```

```r
#Justification for probowl as numeric
xyplot(log~probowl|pos, data=nfl, type=c("p","r"))
```

```r
xyplot(log~heisman|pos, data=nfl)
```

```r
require(dplyr)
nfl <- nfl %>%
  mutate(ivy = ifelse(college %in% c("Boise State", "Boise St", "Oregon", "Mississippi State", "Utah", "Florida State", "Virgini
nfl$ivy <- as.factor(nfl$ivy)
require(ggplot2)
ggplot(nfl, aes(x=ivy, y=salary))+geom_point()+facet_wrap(~exp)
```

```r
nfl.sub <- nfl %>%
  select(pos, xofteams, probowl, ivy, champ, exp, log, team, ht, wt, age, race, hsstate, heisman)
lm1 <- lm(log ~.+pos*ht+pos*wt+ht*wt+age*exp+exp*champ+wt*probowl+team*exp+race*wt+xofteams*probowl, data=nfl.sub)
step.back <- step(lm1, direction="backward")
step.both <- step(lm1, direction="both")

lm0 <- lm(log~exp, data=nfl.sub)
step.for <- step(lm0, direction="forward", scope=list(upper=formula(lm1),lower=~1))
step.for.both <- step(lm0, direction="both", scope=list(upper=formula(lm1),lower=~1))
```

```r
require(nlme)
require(xtable)
fit.test <- lm(log~pos+xofteams+probowl+exp+ht+wt+age+heisman+exp*age, data=nfl.sub)
#extractAIC(fit.test)
xtable(summary(fit.test))
fit.lme <- lme(log~pos+xofteams+probowl+exp+ht+wt+age+heisman+exp*age, data=nfl.sub, random=~1|team, method="ML")#random interce
```

```r
require(xtable)
fit.lme <- lme(log~pos+xofteams+probowl+exp+ht+wt+age+heisman+exp*age, data=nfl.sub, random=~1|team, method="ML")#random interce
print(xtable(ranef(fit.lme), digits=c(0,10)))
```

```
DATA nfl;
  INFILE "/folders/myfolders/nfl.csv" firstobs=2 delimiter =',' dsd;
  INPUT pos $ xofteams probowl ivy $ champ $ exp log team $ ht
   wt age race $ hsstate $ heisman $;
  ;
```

```
RUN;

PROC PRINT DATA=nfl;
RUN;

PROC GLMSELECT data=nfl plot=CriterionPanel;
CLASS pos probowl ivy champ team race hsstate heisman;
MODEL log = pos xofteams probowl ivy champ exp team ht wt age race hsstate
  heisman pos*ht pos*wt ht*wt age*exp exp*champ wt*probowl team*exp race*wt
  xofteams*probowl / selection=BACKWARD stats=all;
RUN;

PROC GLMSELECT data=nfl plot=CriterionPanel;
CLASS pos probowl ivy champ team race hsstate heisman;
MODEL log = pos xofteams probowl ivy champ exp team ht wt age race hsstate
  heisman pos*ht pos*wt ht*wt age*exp exp*champ wt*probowl team*exp race*wt
  xofteams*probowl / selection=FORWARD stats=all;
RUN;

PROC GLMSELECT data=nfl plot=CriterionPanel;
CLASS pos probowl ivy champ team race hsstate heisman;
MODEL log = pos xofteams probowl ivy champ exp team ht wt age race hsstate heisman
  pos*ht pos*wt ht*wt age*exp exp*champ wt*probowl team*exp race*wt xofteams*probowl
  / selection=stepwise(select=AICC) stats=all;
RUN;


PROC MIXED DATA=nfl method=ml;
  CLASS pos probowl ivy champ team race hsstate heisman;
  MODEL log = pos xofteams probowl exp ht wt age heisman exp*age / S;
  random team / s;
RUN;

PROC MIXED DATA=nfl method=ml;
  CLASS pos probowl ivy champ team race hsstate heisman;
  MODEL log = xofteams probowl exp wt ht*wt exp*age / S;
  random team / s;
RUN;
```