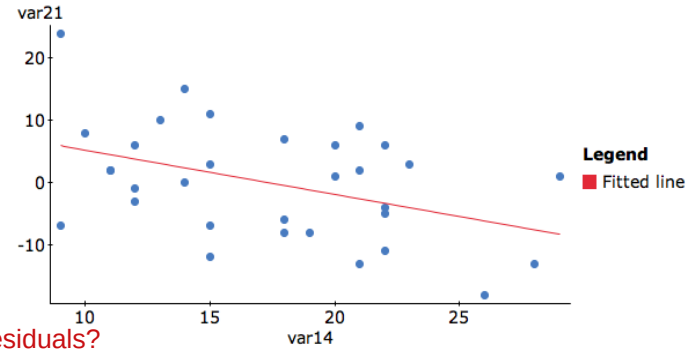
RIT Up-Stat Competition Team: The Outliers

Throughout our research we have found many interesting relationships among the NFL data from the 2009 football season. Not only did we use the data given, but we also used outside data such as subdivision, location (suburban, downtown, or uptown), number or wins and losses, number of draft picks and stadium size.

One interesting and surprising association that we found was between variable 14 and variable 21, which were 4th down attempts and number of turnovers. There was a negative correlation of -0.424, where as the number of 4th down attempts increased the number of turnovers decreased. When there were a negative number of turnovers, it meant that the specific team gave away more turnovers than they received. High 4th down attempts led to more giveaway turnovers then ones received by the teams. The linear model can be expressed as $\widehat{turnovers} = 12.36 - .71\ (4th\ down\ attempts)$
What this means is that for each additional 4th down attempt, the $\widehat{turnovers}$ decreased by about .71 turnovers. Also what this means is that, when the number of 4th downs is zero, the number of turnovers is approximately 12.36.
17.95% of variability in turnovers explains variation in 4th down attempts as well. A linear regression t test can be performed because the data satisfies the following requirements:

- The scatterplot of 4th downs vs. turnovers looks straight enough
- No patterns or fanning in residuals
- Histogram of residuals looks approx. normal

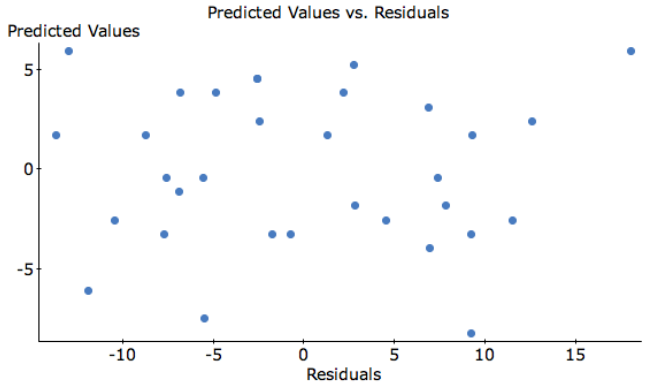


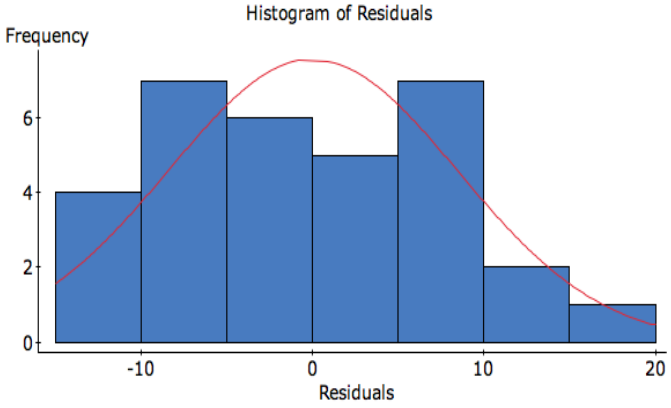
Do you want to put in a picture of the residuals and histogram of residuals?

$H_o = no\ linear\ association\ between\ 4th\ down\ attempts\ and\ tunrovers$
$H_a = there\ is\ a\ linear\ association\ between\ 4th\ down\ attempts\ and\ turovers$
Conclusion: Because my p-value is low, I reject the null. I have evidence that there id a linear association between 4th down attempts and turnovers.

Simple linear regression results:
Dependent Variable: var21 Independent Variable: var14
Sample size: 32 Degrees of Freedom: 30
R (correlation coefficient) = -0.42362503 R-sq. = 0.17945817
Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	12.361447	5.050554	â‰ 0	30	2.4475427	0.0205
Slope	-0.71017287	0.27725041	â‰ 0	30	-2.5614854	0.0157





Also using a linear regression t test, we were able to find an association between the number of losses and number of fumbles. Although this may not be the most interesting or surprising relationship that we found, it does verify our initial thoug of more fumbles leading to losses. This linear model can be used to estimate either the number of losses using fumbles, or fumbles using losses, $\#\ \widehat{losses} = .0586 + .339(fumbles)$

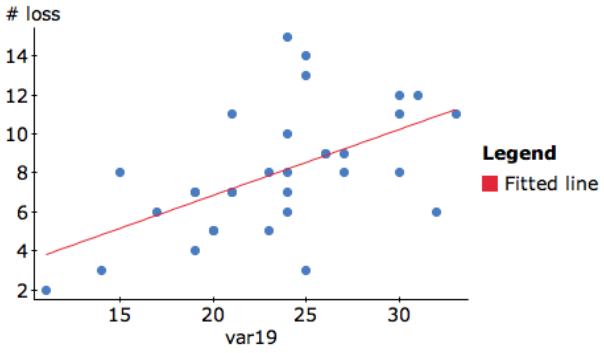
What this means is that for each additional fumbles, the $\#\ of\ \widehat{losses}$ increased by about .339 losses. Also what this means is that, when the number of fumbles is zero, the number of losses is approximately .586. R-sq can be used in the following statement as well, 30.52% of variability in # of losses explains variation in fumbles.

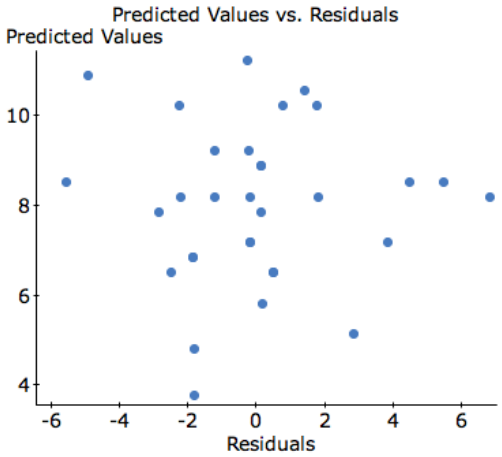
Simple linear regression results:

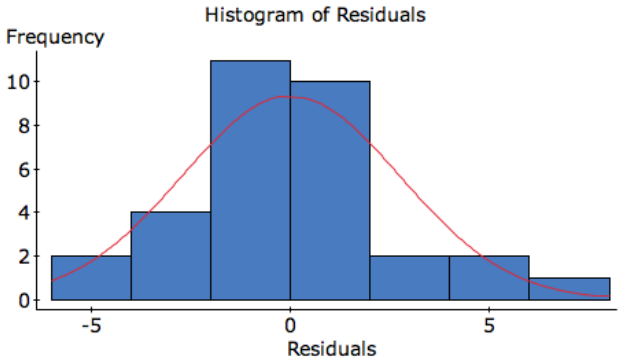
Dependent Variable: # loss — Independent Variable: var19
Sample size: 32 — Degrees of Freedom: 30
R (correlation coefficient) = 0.55241427 — R-sq = 0.30516153
Estimate of error standard deviation: 2.7309216

Parameter estimates:

Parameter	Estimate	Std. Err.	Alternative	DF	T-Stat	P-value
Intercept	0.058565795	2.2404686	≠ 0	30	0.026139976	0.9793
Slope	0.33883453	0.093347795	≠ 0	30	3.6298075	0.001







All of the conditions are satisfied also:

- The scatterplot of losses vs. fumbles looks straight enough
- No patterns or fanning in residuals
- Histogram of residuals looks approx. normal

$H_o = no\ linear\ association\ between\ \#\ if\ losses\ and\ fumbles$
$H_a = there\ is\ a\ linear\ association\ between\ \#\ of\ losses\ and\ fumbles$

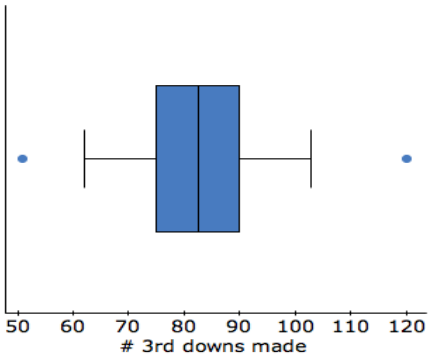
Conclusion: Because my p-value is low, I reject the null. I have evidence that there is a linear association between # losses and fumbles.

I am confused. Before you were looking at bivariate data. Now I am not certain which variable you are discussing.

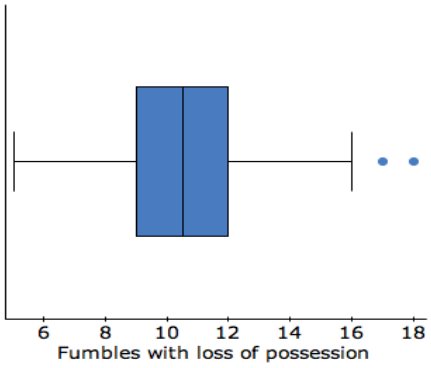
In the given data set there were also some outliers that we found using the outlier test:
$Q_3 + 1.5IQR$ for the upper outliers
$Q_1 - 1.5IQR$ for the lower outliers
How these formulas worked were that we would input the information and get an upper or lower fence or boundary. For the upper boundary, if the max was above it, we knew there had to be at least one upper outlier. Also for the lower boundary, if the min was below it, we knew that there had to be at least one lower outlier.

The outliers that we found included, the number of 3rd downs made throughout the season and the total number of fumbles resulting in a loss of possession during the season.

For the number of 3rd downs, we found the upper outlier to be the Miami Dolphins and the lower outlier to be the Buffalo Bills.



For the fumbles resulting in loss of possession, there were two upper outliers of the Arizona Cardinals and the NY Giants



Using a chi squared independence test, we were also able to conclude that it is mostly likely that the placement of either NFC or AFC teams is random. There is no general area (suburban, downtown, or uptown) that a specific division is placed in.

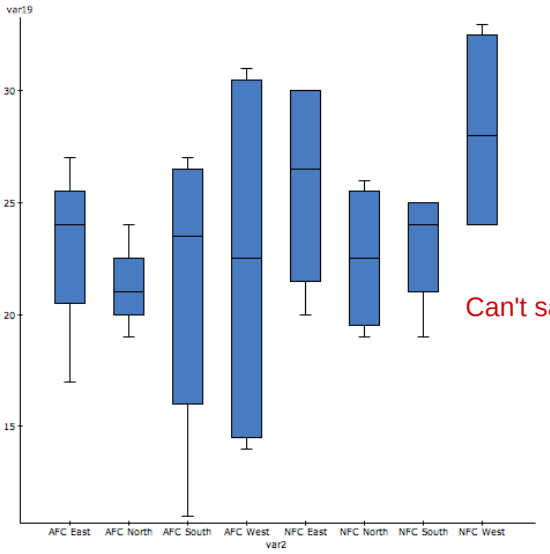
$$H_o: no\ association\ between\ division\ and\ area\ of\ team$$
$$H_a: there\ is\ an\ association\ between\ division\ and\ area\ of\ team$$

Do you want to show all expected values?

	Suburban	Downtown	Intown	Total
NFC	5/5.5	6/5.5	5/5	16
AFC	6	5	5	16
Total	11	11	10	32

C: Counts of division for specific areas of NFL teams
R: not told random; hopefully this season is representative
E: all expected values are greater or equal to 5

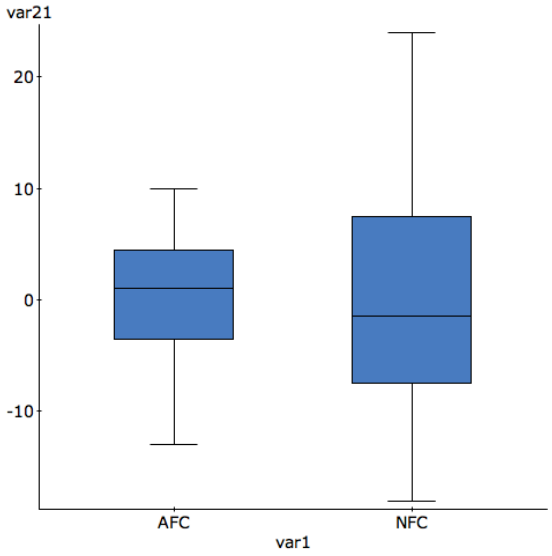
Using these hypothesizes and these conditions, we were able to get a chi-squared value of .181 that then resulted in a p value of about .913.
The conclusion would then be:
Because my p-value is high, I fail to reject the null. I do not have evidence that there is an association between division of teams in the NFL and their location.



This is a boxplot of subdivision vs. number of fumbles. It seems as though the AFC North has the most consistent number of fumbles. They are the most consistent because they have the smallest variation with the smallest range of about 19-24. The AFC South and the AFC West have the same range but the AFC is approximately normal while the AFC South is skewed to the left. The median of the AFC South is about slightly higher than the AFC West and neither have any outliers.

Can't say approx normal with boxplots. Can't see how many modes.

Another interesting thing about this boxplot is that about 100% of the NFC West fumbles are greater than about 75% of the NFC South's fumbles.



It this boxplot of division vs. turnovers, it is very evident that the NFC has a larger variation of turnovers. This can be explained by looking at the lower and upper bounds of each division. The range of the NFC is significantly higher and therefore they are less consistent than the AFC. In addition it could be said that about 50% of the NFC's turnovers are greater than about 75% of the AFC's turnovers. This also supports the previous statement about NFC having a larger variation.