Homework 7

**4.2) Using the teengamb data, fit a model with gamble as the response and the other**

**variables as predictors.**

**(a) Predict the amount that a male with average (given these data) status, income**

**and verbal score would gamble along with an appropriate 95% CI.**

Using average values for each of our predictors – status, income, verbal, and sex – we get an prediction vector:

(Intercept) sex status income verbal

1.000000 0.000000 45.234043 4.641915 6.659574

The sex value was correctly changed to value of 0, corresponding to male. Using this vector to predict how much a male with average predictor values would gamble per year on average, we receive a prediction and confidence interval of:

fit lwr upr

1 28.24252 18.78277 37.70227

Based on this output, a male with predictor values equal to the sample mean for each predictor will gamble $28.24252 per year on average, with the lower band on our confidence interval to be $18.78 per year and the upper band to be $37.70 per year.

**(b) Repeat the prediction for a male with maximal values (for this data) of status,**

**income and verbal score. Which CI is wider and why is this result expected?**

Using the maximum of the sample for each predictor, we get a prediction vector

(Intercept) sex status income verbal

1 0 75 15 10

Using this vector to predict how much a male with max predictor values from our sample would gamble per year, we get a prediction and confidence interval of:

fit lwr upr

1 71.30794 42.23237 100.383

Thus, on average, we would expect a male with predictor values equal to the maximum of our sample to gamble $71.30 per year, with a lower bound on our confidence interval being $42.23 per year, and the upper bound to be $100.38 per year.

**(c) Fit a model with sqrt(gamble)as the response but with the same predictors.**

**Now predict the response and give a 95% prediction interval for the individual**

**in (a). Take care to give your answer in the original units of the response.**

Same we’re using the same predictors, we have a

(Intercept) sex status income verbal

1.000000 0.000000 45.234043 4.641915 6.659574

And we get a prediction and confidence interval of:

fit lwr upr

1 4.049523 3.180676 4.918371

Thus, we get a prediction of

**(d) Repeat the prediction for the model in (c) for a female with status=20,**

**income=1, verbal = 10. Comment on the credibility of the result.**

Using the same model as c), we get a prediction of -$2.086 dolllars per year, which should raise some questions. Specifically, how could an amount of dollars gambled be negative? This doesn’t make much sense; if this were possible, women would certainly be the greatest enemy of all casinos!

**For the fatdata used in this chapter, a smaller model using only age, weight,**

**height and abdom was proposed on the grounds that these predictors are either**

**known by the individual or easily measured.**

**(a) Compare this model to the full thirteen-predictor model used earlier in the**

**chapter. Is it justifiable to use the smaller model?**

After constructing the null and alternative hypothesis, and running the anova() function, we get an output of:

Test statistic: F-stat = 2.9336

P-value = .002

Thus, we have strong evidence to reject the null hypothesis, and claim that given that the age, weight, height, abdom regressors are present in our model, at least one of the other regressors – neck, ankle, biceps, forearm, thigh, hip, wrist, chest, and knee – are not equal to 0 and should be included in our model.

**(b) Compute a 95% prediction interval for median predictor values and compare**

**to the results to the interval for the full model. Do the intervals differ by a**

**practically important amount?**

Using sample median values for our regressors in both models, we get values for prediction and prediction intervals of:

Little Model

fit lwr upr

17.84028 9.696631 25.98392

Big Model

fit lwr upr

17.49322 9.61783 25.36861

Looking at the two intervals, we see negligible changes in their lengths. For the little model, there is a range of 16.28729, and for the big model, we have 15.75078. This is most likely due to the fact that most of the regressors in the big model don’t add much value to our model.

**(c) For the smaller model, examine all the observations from case numbers 25 to**

**50. Which two observations seem particularly anomalous?**

Looking through case numbers 25-50, we can see there are two anomalous values in the model. One corresponds to case 39, which has a weight of 363.15 in pounds, and one corresponds to case 42, which has a height of 29.50 inches.

**(d) Recompute the 95% prediction interval for median predictor values after these**

**two anomalous cases have been excluded from the data. Did this make much**

**difference to the outcome?**

After removing these values, we get the following prediction and prediction interval:

Before

fit lwr upr

17.84028 9.696631 25.98392

After

fit lwr upr

1 17.9033 9.887851 25.91874

And we can see that there wasn’t much change. This should make sense, because we were using the sample **median** values for our predictors, which aren’t biased towards outliers.