1. Fit a model to predict *RPG* with *Games*, *k\_pct*, *BA*, and *bb\_pct*. Write down the model.

```{r}

full <- lm(RPG ~ Games + k\_pct + BA + bb\_pct, data = umpires\_df)

full

```

1. Predict the *RPG* for an umpire that has umpired 255 games, with a *BA* for hitters of 0.270, a *k\_pct* of 15%, and a *bb\_pct* of 10%.

The umpire would average 9.44 runs per game.

1. Obtain the p-values for each of the predictors in the model. Are all of the predictors effective? If not, which ones and why?

```{r}

summary(full)

```

**Games:** 1.61e-05

**k\_pct:** <2e-16

**BA:** <2e-16

**bb\_pct:** 0.000104

All predictors are effective, since they all have p-values less than 5%.

1. Fill in the ANOVA table for the model and perform a test for the overall fit of the model.

```{r}

anova(full)

1-pf(398367.7,4,931)

```

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **d.f.** | **Sum of Squares** | **Mean Square** | **F-value** | **P-value** |
| **Model** | k = **4** | 0.786+10.125+272.38+2.541= **285.832** | SSModel/k = 285.832/4 = **71.458** | MSModel/MSE = 71.458/0.000179377=**398367.7** | **0** |
| **Error** | n-k-1= 936-4-1= **931** | **0.167** | SSError/(n-k-1)= 0.167/931=**0.000179377** |
| **Total** | n-1 = 936-1=**935** | 0.167 + 285.832 = **285.999** |

**Conclusion:** Reject H0; 0<0.05

**Interpretation:** We have significant evidence that the model using *Games, k\_pct, BA,* and *bb\_pct* to predict *RPG* is effective.

1. Record and comment on the multiple R2 and adjusted R2 values for the model.

```{r}

summary(full)

```

Multiple R2: 0.6472

Adjusted R2: 0.6457

The R2 values imply that about 65% of runs per game can be explained by the variables in the model.

1. Drop *bb\_pct* from the model and record the new multiple R2 and adjusted R2 values. Is the model without *bb\_pct* better or worse?

```{r}

new <- lm(RPG ~ Games + k\_pct + BA, data = umpires\_df)

summary(new)

```

Multiple R2: 0.6415

Adjusted R2: 0.6403

Since the R2 values are slightly lower than in the previous model it is better with *bb\_pct* included.

1. Find the VIFs for the original model, are any of the values concerning? Explain why or why not.

```{r}

vif(full)

```

All of the VIFs are below 4 so none of them are of concern.

1. What would it mean if a predictor had a poor VIF value?

This would mean that the predictor had high multicollinearity, meaning some of the predictors are too closely related to one another, this can cause issues with the validity of t.test results.

1. What would you do with a predictor with too high a VIF?

Drop it from the model.

1. Find and interpret the 95% confidence interval for the mean *RPG* for umpires that have umpired 400 games, with a *BA* for hitters of 0.215, a *k\_pct* of 18%, and a *bb\_pct* of 9%.

```{r}

new <- data.frame(Games=400, k\_pct=18, BA=0.215, bb\_pct = 9)

predict(full, new, interval="confidence")

```

(4.933806, 6.573678)

We are 95% confident that the mean *RPG* for all umpires that have umpired 400 games, with a *BA* for hitters of 0.215, a *k\_pct* of 18%, and a *bb\_pct* of 9% will be between 5.59 and 5.92.

1. Find and interpret the 95% prediction interval for the *RPG* of an umpire that has umpired 400 games, with a *BA* for hitters of 0.215, a *k\_pct* of 18%, and a *bb\_pct* of 9%.

```{r}

new <- data.frame(Games=400, k\_pct=18, BA=0.215, bb\_pct = 9)

predict(full, new, interval="prediction")

```

(4.933806, 6.573678)

We are 95% confident that the *RPG* for an umpire that has umpired 400 games, with a *BA* for hitters of 0.215, a *k\_pct* of 18%, and a *bb\_pct* of 9% will be between 4.93 and 6.57.