Econometrics HW #2 Solutions

Ryan Safner

Due: Monday, October 8, 2018

Note: Answers may be longer than I would deem sufficient on an exam. Some might vary slightly based on points of interest, examples, or personal experience. These suggested answers are designed to give you both the answer and a short explanation of *why* it is the answer.

Theory & Concepts

For the following questions, please answer the questions completely but succinctly (2-3 sentences).

1. In your own words, describe what R^2 means. How do we calculate it, what does it tell us, and how do we interpret it?

The R^2 is a measure of how well the OLS regression line "fits" our observed data points. It is a measure of the share of the total variation in Y (TSS) that is explained by the variation from our model (ESS), where:

$$\begin{aligned}
R^2 &= \frac{ESS}{TSS}\\
ESS &= \sum(\hat{Y}_i - \bar{Y})^2\\
TSS &= \sum(Y_i - \bar{Y})^2
\end{aligned}$$

Equivalently, you can also estimate R^2 as:

$$\begin{aligned}
R^2 &= 1 - \frac{SSE}{TSS}\\
SSE &= \sum(\epsilon_i)^2\\
R^2 &= [corr(X,Y)]^2
\end{aligned}$$

The closer R^2 is to 1, the better the fit, the closer to 0, the poorer the fit. Low R^2 tells us that there are better models, including more variables, that explain the variation in Y.

---

2. In your own words, describe what the standard error of the regression (SER) means. How do we calculate it, what does it tell us, and how do we interpret it?

SER ($\hat{\sigma}$) is the average size of the error (a.ka. the residual), that is, the average distance from the regression line to the actual data value for Y at a given X. The goal of OLS is to minimize this (well, technically just the SSE).

$$
\begin{aligned}
SER &= \sqrt{\frac{1}{n-2}\sum \hat{\epsilon_i}^2} \\
&= \sqrt{\frac{SSE}{n-2}}
\end{aligned}
$$

We calculate it by squaring the residuals (to get a positive distance), taking the mean of them by adding them all up and dividing by $n-2$, and then taking the square root to return to normal (non-squared) units. We divide by $n-2$ rather than n due to the correction for calculating two parameters with our data already, $\hat{\beta_0}$ and $\hat{\beta_1}$.

---

3. In your own words, describe what exogeneity and endogeneity mean, and how they are related to bias. What can we learn about the bias if we know X is endogenous?

The OLS estimators $\hat{\beta_0}$ and $\hat{\beta_1}$ are unbiased estimates of the true population parameters β_0 and β_1 if and only if X is *exogenous.* That is to say, if $corr(X, \epsilon) = 0$ (i.e. there is no correlation between X and any unobserved variable that affects Y), then $E[\hat{\beta_1}] = \beta_1$.

If X *is* correlated with the error term, then X is *endogenous.* The true expected value of the OLS estimator is

$$E[\hat{\beta_1}] = \beta_1 + corr(X, \epsilon)\frac{\sigma_\epsilon}{\sigma_X}$$

The bias is $\big(E[\hat{\beta_1}] - \beta_1\big)$, i.e. the difference between average estimated sample slope and the 'true' population slope, so we can determine first the *size* of the bias based on how large $corr(X, \epsilon)$ is. The stronger the correlation, the larger the bias.

Second, we can determine the *direction* of the bias depending on the sign of $corr(X, \epsilon)$.

- If X and ϵ are positively correlated (move in the same direction), we know that we have *overstated* the true effect of ΔX on ΔY, since a change in Y is picking up both a change in X and a further change (in the same direction as X) in the unobserved ϵ.
- If the correlation is negative (move in opposite directions), we know that we have *understated* the true effect of ΔX on ΔY, since a change in Y is picking up both a change in X that is dampened by a change in the opposite direction of ϵ.

---

4. In your own words, describe what homoskedasticity and heteroskedasticity mean: both in ordinary English, and in terms of the graph of the OLS regression line.

Homoskedasticity means the errors are distributed with the same variance for all levels of X. Knowing anything about X will not tell us anything about the distribution of errors at that level of X.

Heteroskedasticity means the errors are distributed differently for different levels of X. So, at different levels of X, there will be much more or much less variation in the residuals.

---

5. A researcher is interested in examining the impact of illegal music downloads on commercial music sales. The author collects data on commercial sales of the top 500 singles from 2017 (Y) and the number of downloads from a web site that allows 'file sharing' (X). The author estimates the following model

$$\text{music sales}_i = \beta_0 + \beta_1 \text{illegal downloads}_i + \epsilon_i$$

The author finds a large, positive, and statistically significant estimate of $\hat{\beta_1}$. The author concludes these results demonstrate that illegal downloads actually *boost* music sales. Is this an unbiased estimate of the impact of illegal music on sales? Why or why not? Do you expect the estimate to overstate or understate the true relationship between illegal downloads and sales?

Does knowing the amount of illegal downloads an artist has convey any information about other variables that affect music sales? In other words, we are asking if $E[\epsilon|X] = 0$ (or more simply, $corr(X, \epsilon) = 0$).

It is likely that artists and songs that are the most heavily pirated are the most popular ones, and also are likely have very high music sales. Economists say piracy is like a tax on success–it happens more to those who are already successful and less to those who are still trying to make it big.

In any case, illegal downloads is probably endogenous. Since there is likely a positive correlation between music sales and popularity (in the error term), and popularity is also positively correlated with music sales, it is likely that we are *overstating* the effect of illegal downloads on sales. In other words, $\hat{\beta_1}$ is also picking up the positive effect of popular songs, and is too large. The true estimate of β_1 is likely much lower than measured.

---

6. A pharmaceutical company is interested in estimating the impact of a new drug on cholesterol levels. They enroll 200 people in a clinical trial. People are randomly assigned the treatment group or into the control group. Half of the people are given the new drug and half the people are given a sugar pill with no active ingredient. To examine the impact of dosage on reductions in cholesterol levels, the authors of the study regress the following model:

$$\text{cholesterol level}_i = \beta_0 + \beta_1 \text{dosage level}_i + \epsilon_i$$

For people in the control group, dosage level$_i$ = 0 and for people in the treatment group, dosage level$_i$ measures milligrams of the active ingredient. In this case, the authors find a large, negative, statistically significant estimate of $\hat{\beta_1}$. Is this an unbiased estimate of the impact of dosage on change in cholesterol level? Why or why not? Do you expect the estimate to overstate or understate the true relationship between dosage and cholesterol level?

Does knowing whether (or how much) a person was treated convey any information about other characteristics that affect cholesterol level (in ϵ_i)? Again, we are asking if $E[\epsilon|X] = 0$ or $corr(X, \epsilon) = 0$

```
In this case, the answer is clearly no, the equations do hold and treatment is exogenous. $X_i$ is dete
```

Theory Problems

For the following questions, please *show all work* and explain answers as necessary. You may lose points if you only write the correct answer. You may use *R* to verify your answers, but you are expected to reach the answers in this section "manually."

7. Suppose a researcher, using data on class size and average test score from 100 classes, estimates the following OLS regression:

$$\widehat{\text{Test score}} = 520.4 - 5.82\text{Class size}, R^2 = 0.08, SER = 11.5$$

a. Interpret what $\hat{\beta_0}$ means in this context.

$\hat{\beta_0}$ is the intercept, literally the average test score for a class size of 0 would be 520.4.

b. Interpret what $\hat{\beta_1}$ means in this context.

$\hat{\beta_1}$ is the slope, $\frac{\Delta\text{Test Score}}{\Delta\text{Class Size}}$ implying that for every 1 student increase (decrease) in class size, average test score will change by -5.82 (+5.82).

c. A class has 22 students. What is the regression's prediction for that classroom's average test score?

$$\widehat{\text{Test score}} = 520.4 - 5.82(22) = 392.36$$

d. It turns out the class with 22 students had an actual average test score of 401. What is the residual for this class?

$$\begin{aligned} \hat{\epsilon_{22}} &= \widehat{\text{Test score}}_{22} - \text{Test score}_{22} \\ &= 401 - 392.36 \\ &= 8.64 \end{aligned}$$

8. A researcher wants to estimate the relationship between average weekly earnings (AWE, measured in dollars) and age (measured in years) using a simple OLS model. Using a random sample of college-educated full-time workers aged 25-65 yields the following:

$$\widehat{AWE} = 696.7 + 9.6 \times Age, R^2 = 0.023, SER = 624.1$$

a. Interpret what the coefficients 696.7 and 9.6 mean.

696.7 is the intercept of the OLS regression line, where it crosses the vertical-axis. It literally means what is the average earnings for a person age 0 years, and as such is of no practical value here.

9.6 is the slope of the regression line, and implies that for every additional year older a person is, their average weekly earnings increase by $9.60.

---

b. What are the units of the SER in this context, and what does it mean? Is the SER large in the context of this regression?

SER is measured in the same units as the dependent variable, AWE, so it is measured in dollars. It is the average error or residual for an individual, the difference (in dollars) between OLS' predicted AWE for that person, and their true AWE in the data. This SER is quite big, $624 in average weekly earnings.

---

c. The R^2 for the regression is 0.023. What are the units of the R^2, and what does it mean?

R^2 is unitless, and measures the "goodness of fit," of the regression - technically, how much the total variance in AWE is measured by the variance in Age, according to our model ($\frac{ESS}{TSS}$)

---

d. What does the regression predict will be the earnings of a 25 year-old worker? How about a 45 year-old worker?

$$\widehat{AWE}_25 = 696.7 + 9.6(25) = \$936.70$$

$$\widehat{AWE}_45 = 696.7 + 9.6(45) = \$1128.70$$

---

e. What does the error term (ϵ_i) represent in this case, and why might individuals have different values of ϵ_i?

The error term represents factors other than age that affects an individual's average weekly earnings. This could include things like experience, ability, job type, education level, etc.

---

f. Do you think it's likely that age is exogenous? Why or why not? Would we expect $\hat{\beta_1}$ to be too large or too small?

Very unlikely. Knowing someone's age likely gives us information about ϵ: we can guess about their experience or level of education (they are likely higher for older people), and these positively affect wages. Thus, we have probably *overstimated* the effect of age on earnings (i.e. $\hat{\beta_1}$), and the true β_1 is likely smaller.

---

9. Suppose a researcher is interested in estimating the linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

and in a sample of 48 observations, generates the following descriptive statistics:

- $\bar{X} = 30$
- $\bar{Y} = 63$
- $\displaystyle\sum_{i=1}^{n}(X_i - \bar{X})^2 = 6900$
- $\displaystyle\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 29000$
- $\displaystyle\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = 13800$
- $\displaystyle\sum_{i=1}^{n}\hat{\epsilon}^2 = 1656$

a. What is the OLS estimate of $\hat{\beta_1}$?

The formula for $\hat{\beta_1} = \dfrac{\displaystyle\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\displaystyle\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{cov(X,Y)}{var(X)} = \frac{13800}{6900} = 2$

---

b. What is the OLS estimate of $\hat{\beta_0}$?

The formula for $\hat{\beta_0} = \bar{Y} - \hat{\beta_1}\bar{X} = 63 - 30(2) = 3$

---

c. Suppose the OLS estimate of $\hat{\beta_1}$ has a standard error of 0.072. Without running a t-test, could we probably reject a null hypothesis of $H_0 : \beta_1 = 0$ at the 95% level?

Yes, we could reject the null hypothesis as the estimate of $\hat{\beta_1} = 2$ is more than 1.96 times its standard error of 0.072. This would generate a t-score higher than 1.96, making it beyond the critical value needed to reject H_0.

---

d. Calculate the R^2 for this model. Does this model explain a lot of variation in Y_i?

We know TSS (4th bullet point) and SSE (last bullet point).

$$\begin{aligned} R^2 &= 1 - \frac{SSE}{TSS} \\ &= 1 - \frac{1656}{29000} \\ &= 1 - 0.057 \\ &= 0.943 \end{aligned}$$

Yes this model explains 94.3% of the variation in Y_i.

---

e. How large is the average residual?

We need to find the standard error of the regression, but luckily we know the SSE (last bullet point)

$$\begin{aligned} SER &= \sqrt{\frac{SSE}{n-2}} \\ &= \sqrt{\frac{1656}{48-2}} \\ &= \sqrt{36} \\ &= 6 \end{aligned}$$

This tells us the average residual is 36 (units of Y).

R Problems

For the following problems, please attach/write the answers to each question on the same document as the previous problems, but also include a printed/attached (and commented!) *.R* script file of your commands to answer the questions.

10. Download the `MLBattend` dataset from Blackboard. This data contains data on attendance at major league baseball games for all 32 teams from the 1970s-2000. Edit the following commands to import the data into an object called `MLBattend`.

```
# install.packages("foreign") # if you don't have it installed, to load .dta file
library("foreign") # load foreign
# MLBattend<-read.dta(/path/to/downloaded/file) # edit to where you downloaded MLBattend.dta
# e.g. for me it's
MLBattend<-read.dta("~/Dropbox/Teaching/Hood College/ECON 480 - Econometrics/Data/MLBattend.dta") #comm
```

a. Get summary statistics for `home_attend` and `runs_scored`

```
summary(MLBattend$home_attend)
```

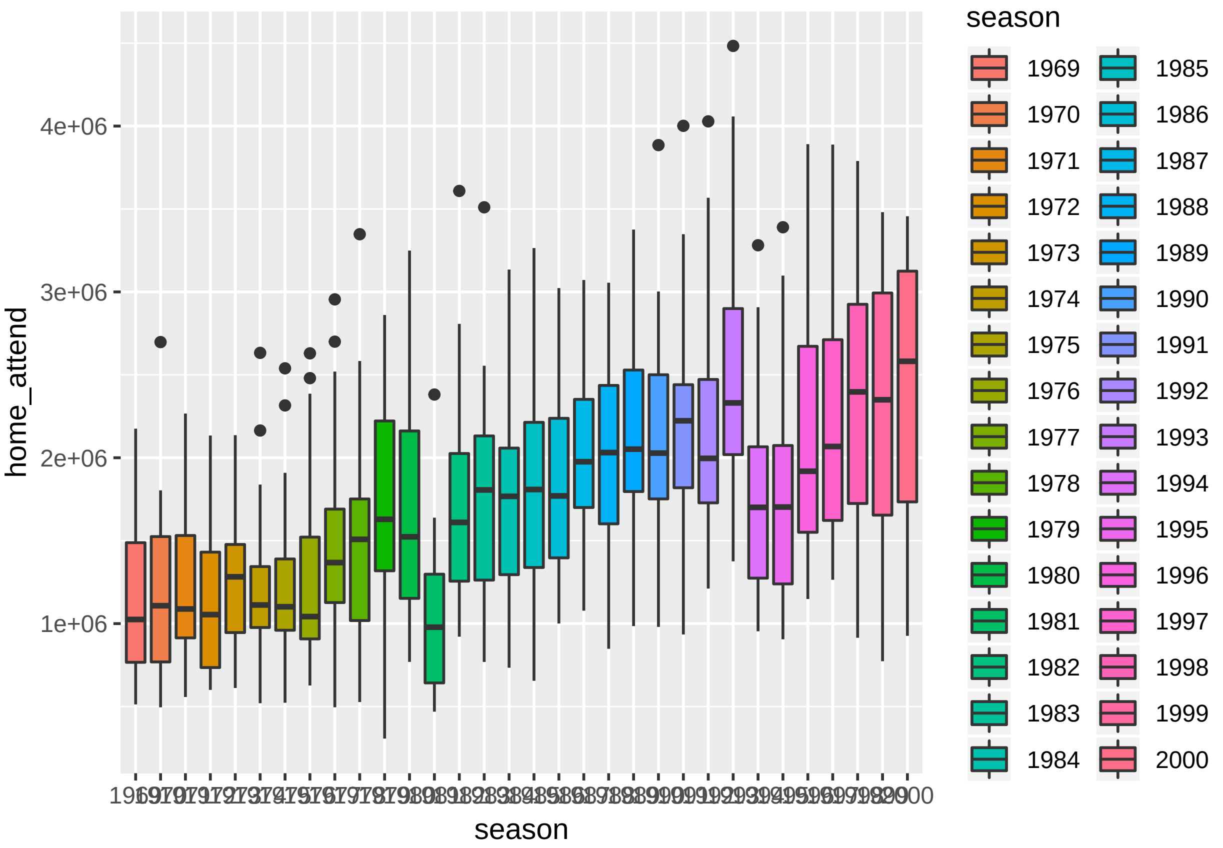
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  306763 1195865 1681896 1777994 2278022 4483350
```

```
summary(MLBattend$runs_scored)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   329.0   633.0   691.5   694.9   759.0  1009.0
```

b. Create a boxplot for `home_attend` over time (that is, over the `season`s). In order to do this, redefine `season` as a factor with `as.factor()` (so `R` knows to use season as a categorical variable). How does attendance seem to change over time?

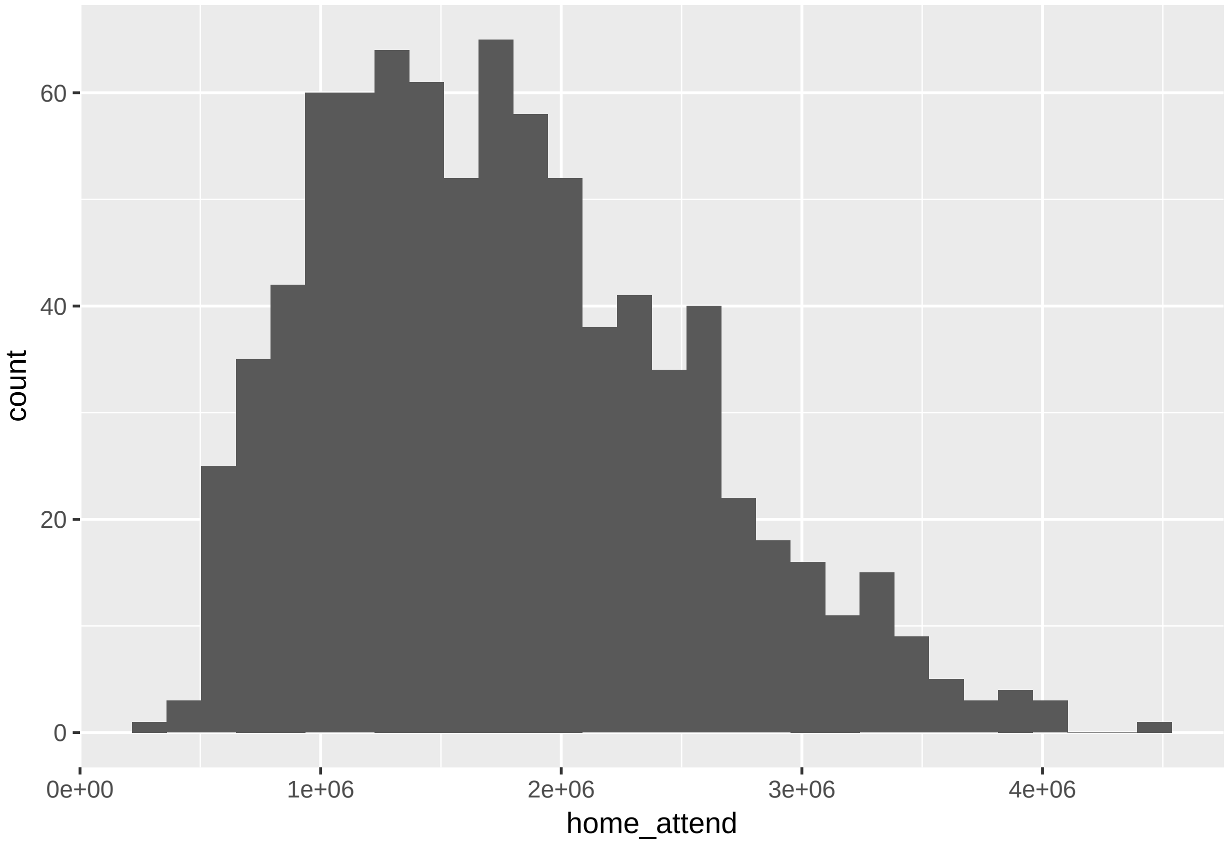
```
library("ggplot2")
MLBattend$season<-as.factor(MLBattend$season) # reclass into factor
p<-ggplot(MLBattend,aes(x=season,y=home_attend,fill=season))+
  geom_boxplot()
p
```



c. Create two histograms (each in percents), one for `home_attend` and one for `runs_scored`. Describe the skew of each distribution, and why this makes sense.

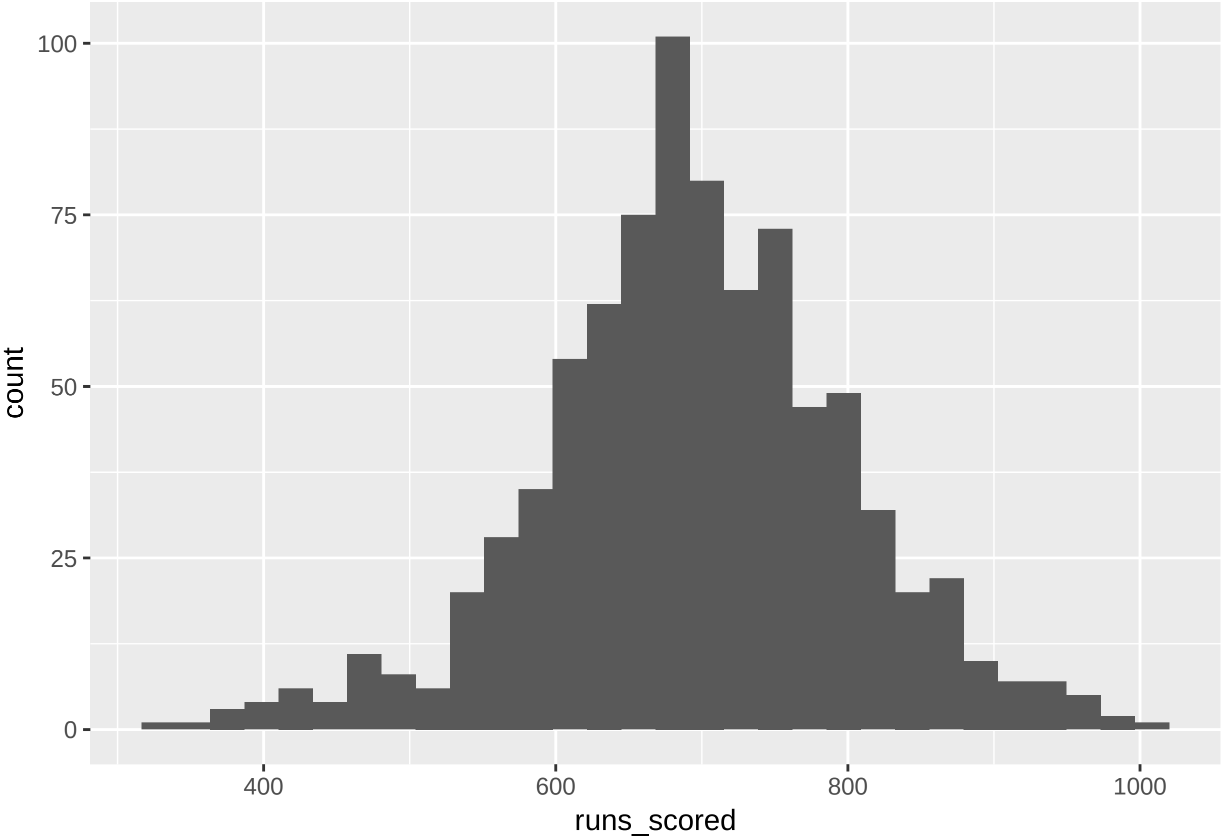
```
h.attend<-ggplot(MLBattend,aes(x=home_attend))+
  geom_histogram()
h.attend
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
h.runs<-ggplot(MLBattend,aes(x=runs_scored))+
  geom_histogram()
h.runs
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

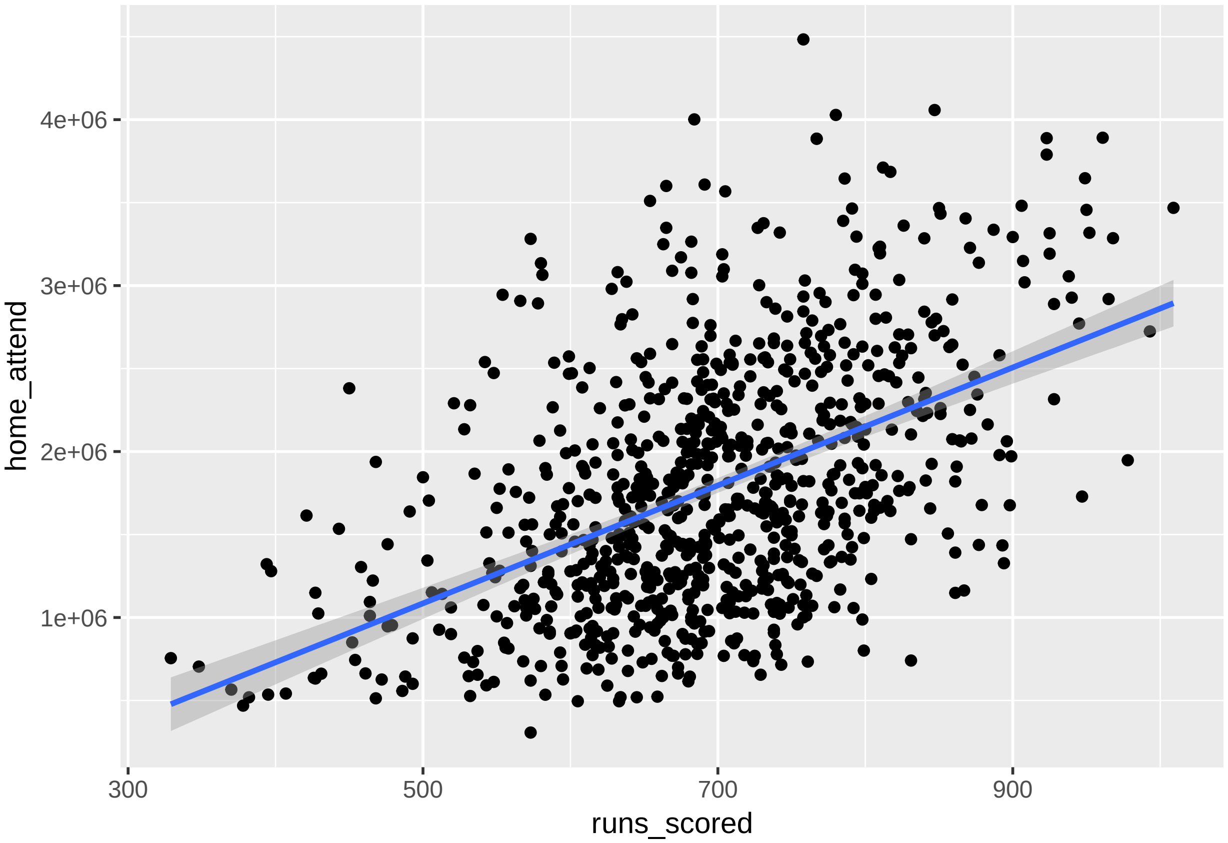


Attendance is skewed to the right. Some teams probably are from larger cities or have larger fanbases than less popular teams, so there is probably larger variation (or at least some outliers) at the top end of the distribution. Runs are approximately normal and symmetric.

d. Create a scatterplot between `home_attend` (as dependent variable) and `runs_scored` (as independent variable). Add a regression line to the scatterplot.

```
scatter<-ggplot(MLBattend, aes(x=runs_scored,y=home_attend))+
  geom_point()+
  geom_smooth(method="lm")

scatter
```



e. Estimate the following regression model:

$$\text{Home attendance rate}_i = \beta_0 + \beta_1 \text{runs scored}_i + \epsilon_i$$

Write the equation of the regression, placing standard errors in parentheses beneath the coefficients. Round to the nearest whole number. Assume the errors are homoskedastic. Then interpret each coefficient. Finally, can we reject the null hypothesis at the 5% level that there is no relationship between runs and attendance?

```
reg<-lm(home_attend~runs_scored, data=MLBattend)

summary(reg)
```

```
##
## Call:
## lm(formula = home_attend ~ runs_scored, data = MLBattend)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1520764  -477972   -84081   422871  2481272
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -691487.6   151852.8  -4.554 6.06e-06 ***
## runs_scored    3553.5      216.1  16.447  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 657400 on 836 degrees of freedom
## Multiple R-squared:  0.2445, Adjusted R-squared:  0.2436
## F-statistic: 270.5 on 1 and 836 DF,  p-value: < 2.2e-16
```

$$\widehat{\text{Home Attendance}} = \underset{(151853)}{-691488} + \underset{(216)}{3554}\,\text{Runs Scored}$$

For every additional run per year, average attendance will increase by 3,554 people. For years that teams have no runs, average attendance for that team is -691,488. Again, we see the intercept is not meaningful. First, no team has failed to get any runs in a whole season. Second, there can't be negative numbers of people attending.

The null hypothesis that $\beta_1 = 0$ can be safely rejected at the 5% level, as we have a very large t of 16.45, and a p-value of practically 0. Note this means that the estimate for the slope, 3554, is 17 times its standard error of 216.

f. Predict the attendance for a team that scores 500 runs in a year. Also predict the residual(s) for having 500 runs in a year. Is the residual larger or small than the average?

```
# a few ways to do this

# 1) use equation to predict value

runsscored<-500 # define runs value

predict.500<-reg$coefficients[1]+reg$coefficients[2]*runsscored # take coef 1 (beta 0) and 2 (beta 1) f
predict.500 # the predicted attendance for 500 runs
```

```
## (Intercept)
##     1085271
```

```
# 2) if we have an observation of 500 runs in our data

# list the predicted value ("fitted.value") from regression for runs equal to 500
predict.500<-reg$fitted.values[MLBattend$runs_scored==500]
predict.500
```

```
##     189
## 1085271
```

```
# list the actual value of attendance for runs of 500
actual.500<-MLBattend$home_attend[MLBattend$runs_scored==500]
actual.500
```

```
## [1] 1845208
```

```
# find residual
res.500<-actual.500-predict.500
res.500
```

```
##      189
## 759937.5
```

```
# alternatively

res.500<-reg$residuals[MLBattend$runs_scored==500]
res.500
```

```
##      189
## 759937.5
```

The residual is 759,938. The *average* residual is given by the standard error of the regression, which is 657,400. This residual for the team with 500 runs is thus larger than the average residual.

g. Look at some other variables that might affect attendance, and present them in a nice table. Run separate regressions of attendance on runs allowed, wins, losses, and games behind. Present them in a nice table using stargazer.

```
reg2<-lm(home_attend~runs_allowed,data=MLBattend)
reg3<-lm(home_attend~wins,data=MLBattend)
reg4<-lm(home_attend~losses,data=MLBattend)
reg5<-lm(home_attend~games_behind, data=MLBattend)

library("stargazer")
```

```
##
## Please cite as:
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
stargazer(reg,reg2,reg3,reg4,reg5, type="latex")
```

```
##
## % Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harva
## % Date and time: Tue, Oct 02, 2018 - 21:23:27
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{5}{c}{\textit{Dependent variable:}} \\
## \cline{2-6}
## \\[-1.8ex] & \multicolumn{5}{c}{home\_attend} \\
## \\[-1.8ex] & (1) & (2) & (3) & (4) & (5)\\
## \hline \\[-1.8ex]
##  runs\_scored & 3,553.516$^{***}$ &  &  &  &  \\
##   & (216.055) &  &  &  &  \\
##   & & & & & \\
##  runs\_allowed &  & 680.278$^{***}$ &  &  &  \\
##   &  & (246.618) &  &  &  \\
##   & & & & & \\
##  wins &  &  & 27,345.180$^{***}$ &  &  \\
##   &  &  & (1,833.210) &  &  \\
##   & & & & & \\
##  losses &  &  &  & $-$17,841.210$^{***}$ &  \\
##   &  &  &  & (1,972.693) &  \\
##   & & & & & \\
##  games\_behind &  &  &  &  & $-$26,308.980$^{***}$ \\
##   &  &  &  &  & (2,030.060) \\
##   & & & & & \\
##  Constant & $-$691,487.600$^{***}$ & 1,305,275.000$^{***}$ & $-$378,163.700$^{***}$ & 3,185,235.000$
##   & (151,852.800) & (173,335.400) & (146,400.900) & (157,583.200) & (37,704.130) \\
##   & & & & & \\
## \hline \\[-1.8ex]
## Observations & 838 & 838 & 838 & 838 & 838 \\
## R$^{2}$ & 0.244 & 0.009 & 0.210 & 0.089 & 0.167 \\
## Adjusted R$^{2}$ & 0.244 & 0.008 & 0.209 & 0.088 & 0.166 \\
## Residual Std. Error (df = 836) & 657,405.000 & 752,906.200 & 672,148.600 & 721,835.900 & 690,167.600
## F Statistic (df = 1; 836) & 270.514$^{***}$ & 7.609$^{***}$ & 222.504$^{***}$ & 81.796$^{***}$ & 167
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{5}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

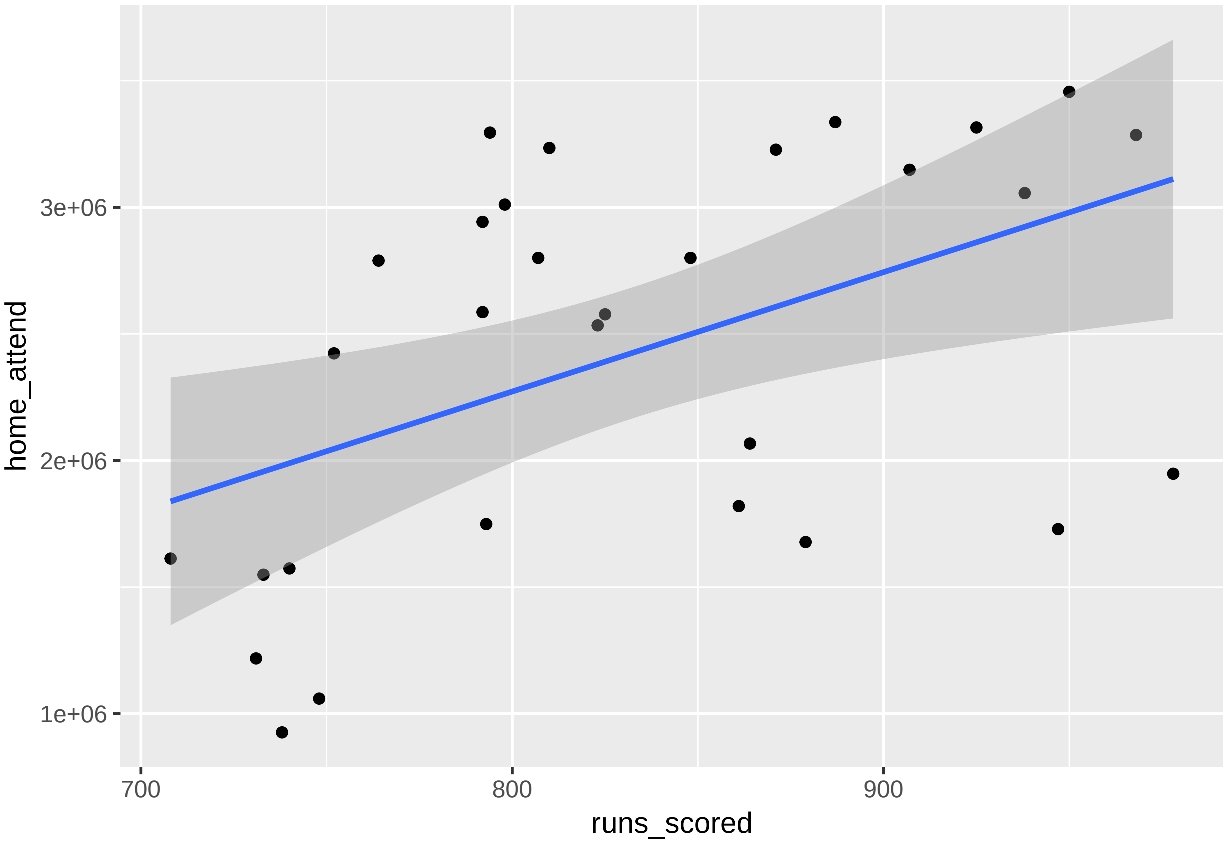
h. Let's look only at the 2000 season. Run a regression and plot a scatterplot with OLS line between attendance and runs scored, but only for the year 2000 (hint: create a new data frame with the `subset()` function, taking `MLBattend` for `season==2000`)

```
MLB2000<-subset(MLBattend, season==2000)

reg2000<-lm(home_attend~runs_scored,data=MLB2000)
summary(reg2000)
```

```
##
## Call:
## lm(formula = home_attend ~ runs_scored, data = MLB2000)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1236607  -502970   204157   490309  1051072
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1499887    1361391  -1.102  0.27996
## runs_scored     4715       1628   2.895  0.00726 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 692200 on 28 degrees of freedom
## Multiple R-squared:  0.2304, Adjusted R-squared:  0.2029
## F-statistic: 8.384 on 1 and 28 DF,  p-value: 0.007263
```

```
scatter.2000<-ggplot(MLB2000,aes(x=runs_scored,y=home_attend))+
  geom_point()+
  geom_smooth(method="lm")
scatter.2000
```



home_attend
3e+06
2e+06
1e+06
700
800
900
runs_scored