Anagha Uppal - MW SECTION - Homework 5

Simple Linear Regression 2

due Sept 29 by 330

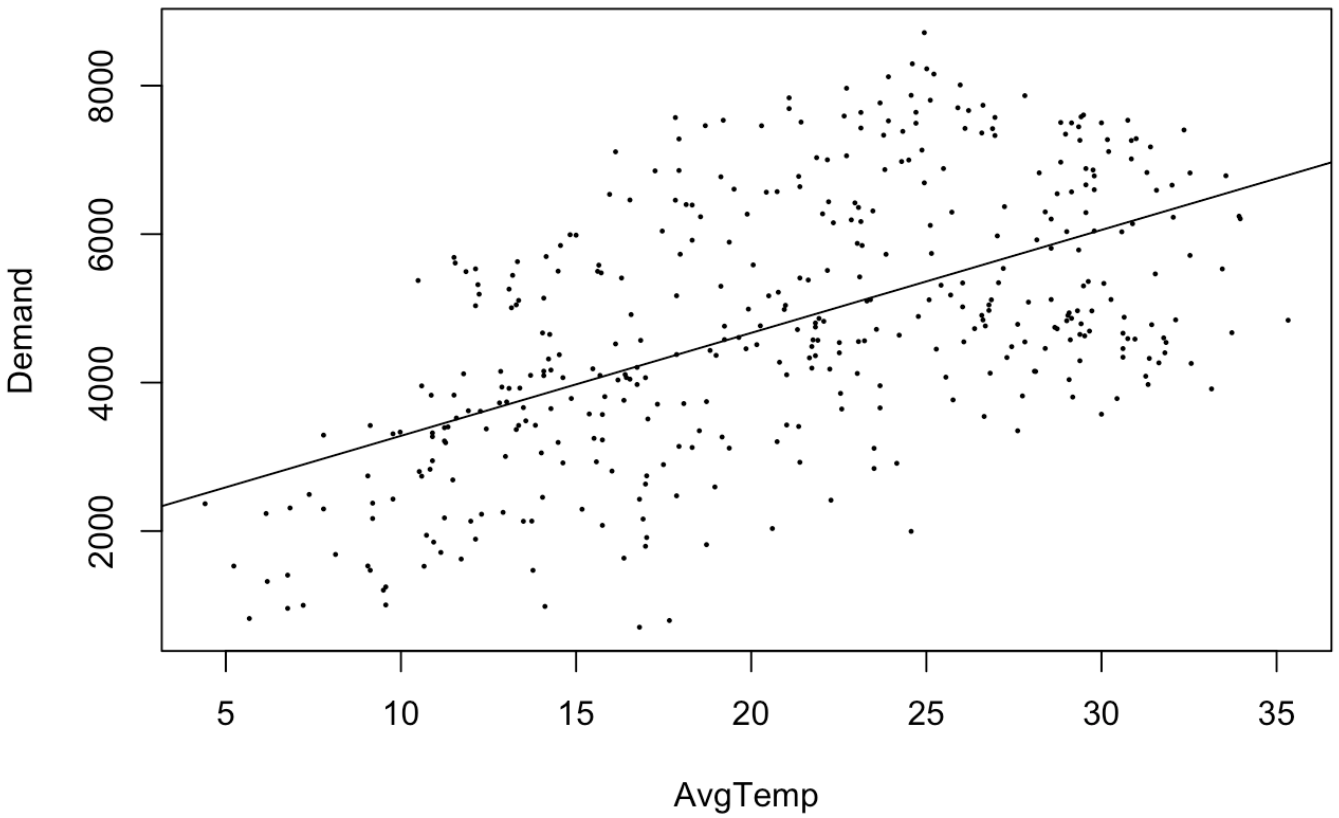
Question 1

Use the `data` command to load in the `EX7.BIKE` dataset. This is the “bike demand” dataset we have used before. Previously, we saw that `AvgTemp` seemed to be the best predictor of `Demand` .

```
#code reading in the BULLDOZER dataset
data("EX7.BIKE")
?EX7.BIKE
```

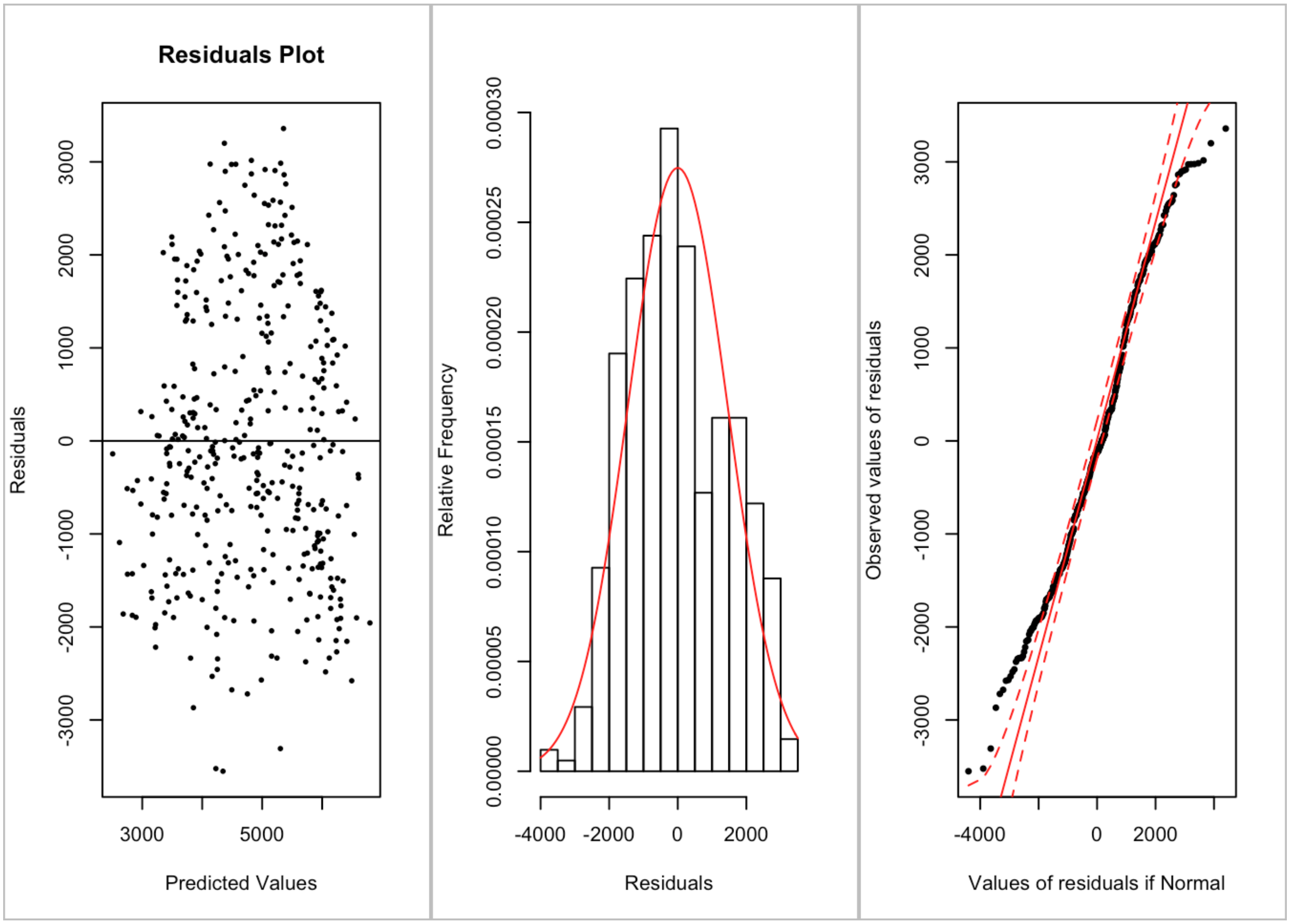
a. Make the scatterplot (add the arguments `pch=20` and `cex=0.3` to make the points small), fit the regression, then add the regression line.

```
#code for plot(), with additional arguments pch=20 and cex=0.3
plot(Demand~AvgTemp,data=EX7.BIKE,pch=20,cex=0.3)

#code fitting and naming the regression model
M <- lm(Demand~AvgTemp,data=EX7.BIKE)

#code adding the regression line
abline(M)
```

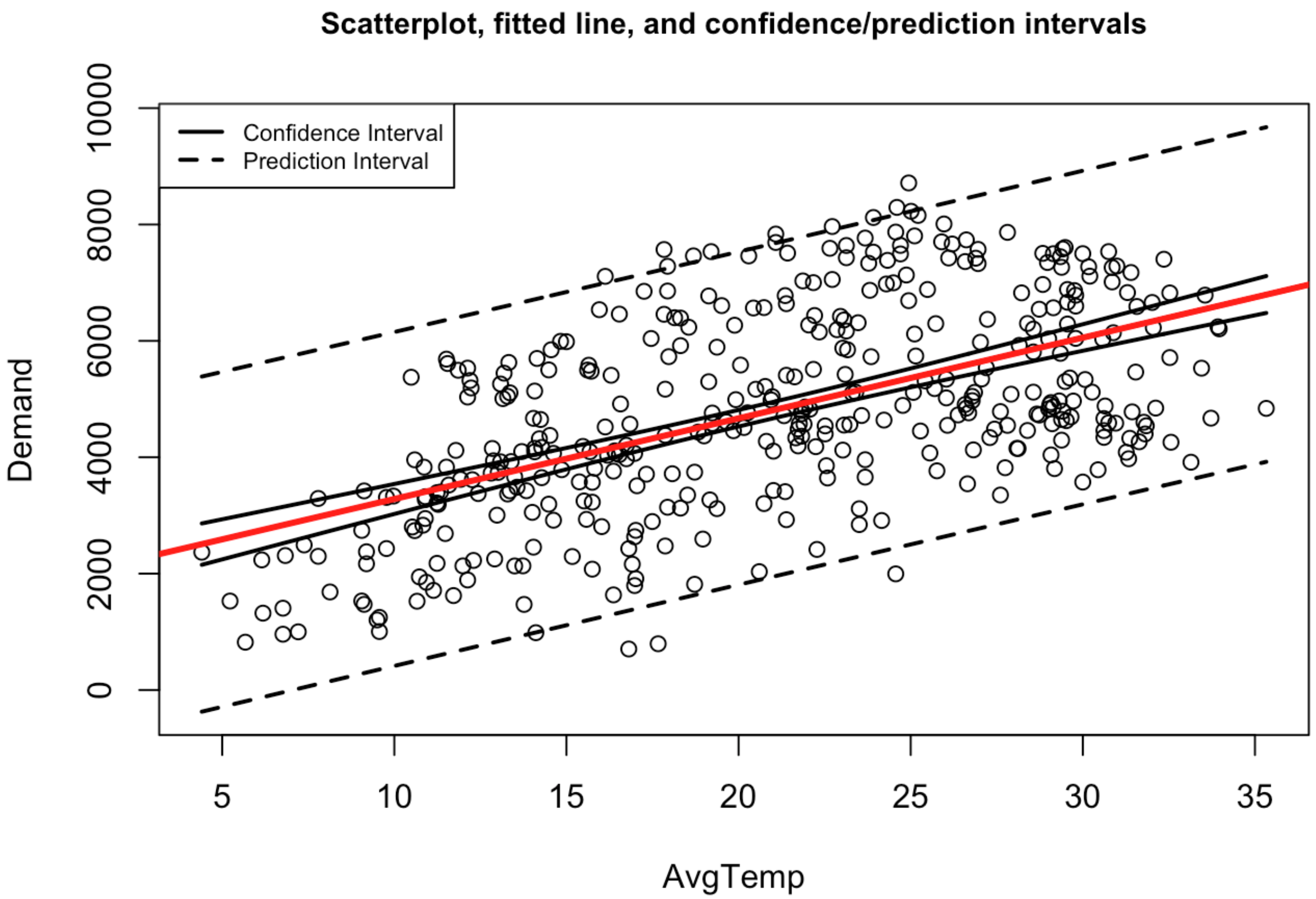


b. Run `check.regression` on the model.

```
#Code running check.regression
check.regression(M)
```



```
##
## Tests of Assumptions: ( sample size n = 410 ):
## Linearity
##    p-value for AvgTemp : 0.0586
##    p-value for overall model : 0.0586
## Equal Spread:  p-value is 0.3376
## Normality:  p-value is 1e-04
##
## Advice:  if n<25 then all tests must be passed.
## If n >= 25 and test is failed, refer to diagnostic plot to see if violation is
severe
##  or is small enough to be ignored.
visualize.model(M)
```



- Is the statistical test for linearity passed? Explain.

Comment: Linearity is passed - p-value is > 5%

- Is the statistical test for equal vertical spread passed? Explain.

Comment: Equal spread is passed - p-value is > 5%

- Is the statistical test for Normality passed? Explain.

Comment: Normality is failed. p-value is too small.

- For the test(s) that were failed, comment on the appropriate diagnostic plot as to whether the violation is severe or relatively minor.

Comment: The violation of the normality does not look severe enough to invalidate the model. There are a few outliers on either end that pull the points outside the red bands, but generally, they seem to stay inside.

- What is your opinion about the model? Does the model reflect reality “close enough” to be used and interpreted or is it useless?

Comment: The sample size is large enough that some differences do not matter too much. Generally, the model seems to be usable. Indeed, over 95% of the points fall inside the prediction interval.

c. Regardless of your assessment, let us interpret what the model says about the relationship for practice. Run `summary` on the model.

```
#Code for summary
summary(M)
```

```
##
## Call:
## lm(formula = Demand ~ AvgTemp, data = EX7.BIKE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3551.3 -1127.5  -118.2  1159.2  3358.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1895.84     222.07   8.537 2.75e-16 ***
## AvgTemp       138.73      10.04  13.819  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1454 on 408 degrees of freedom
## Multiple R-squared:  0.3188, Adjusted R-squared:  0.3172
## F-statistic:   191 on 1 and 408 DF,  p-value: < 2.2e-16
```

```
sd(EX7.BIKE$Demand)
```

```
## [1] 1759.096
```

- Give a detailed interpretation of the slope (rounded to the nearest integer) of the regression line.

Comment With every per degree increase in average temperature, the predicted demand for bikes increases by 139 bikes. The model predicts that at a temperature of 0 degrees, there will be a demand for 1896 bikes.

- Give a detailed interpretation of the model's R^2, rounded to the nearest percent.

Comment 32% of the variation in demand can be attributed to this model (SSE is reduced by 32% when compared with naive model to make predictions)

- Give a detailed interpretation of the model's RMSE.

Comment As compared with the standard deviation (1759.096), the model's RMSE of 1454 is lower, and this is therefore a significantly better model.

---

d. Let us examine the predicted the values of `Demand` when `AvgTemp` equals 10, 20, 30, and 40 (which corresponds to 50F, 68F, 86F, and 104F).

```
#Code defining the data frame used to make predictions
TO.PREDICT <- data.frame( AvgTemp=c(10,20,30,40) )

#Code running predict that obtains confidence intervals
predict(M,newdata=TO.PREDICT,interval="confidence")
##        fit      lwr      upr
## 1 3283.110 3025.298 3540.923
## 2 4670.385 4528.066 4812.704
## 3 6057.660 5829.774 6285.545
## 4 7444.934 7043.066 7846.802

#Code running predict that obtains prediction intervals
predict(M,newdata=TO.PREDICT,interval="prediction")
##        fit       lwr       upr
## 1 3283.110  413.9875  6152.233
## 2 4670.385 1809.3267  7531.443
## 3 6057.660 3191.0707  8924.248
## 4 7444.934 4559.2976 10330.570
```

- Give the point estimate and 95% confidence interval for the average demand on days where `AvgTemp` is 20. Round to the nearest integer.

Comment: Estimated value: 4670. 95% confidence interval is between 4528.066 and 4812.704

- Give the point estimate and 95% prediction interval for tomorrow’s demand (which is expected to have `AvgTemp` of 30). Round to the nearest integer.

Comment: Estimated value: 6057.660. 95% prediction value is between 3191.0707 and 8924.248

- Making a prediction at one of these four temperatures requires extrapolation. Which one? Why is making a prediction at this temperature a bad/dangerous idea?

Comment: The highest AvgTemp value for which data is known is 35.33. Estimating the demand for temperatures higher than that known value requires extrapolation. Extrapolation can at times be dangerous because we cannot be sure that the trend will continue past the listed values in the table. It is not necessary that above a 100 degrees Fahrenheit, the same linear model will apply.

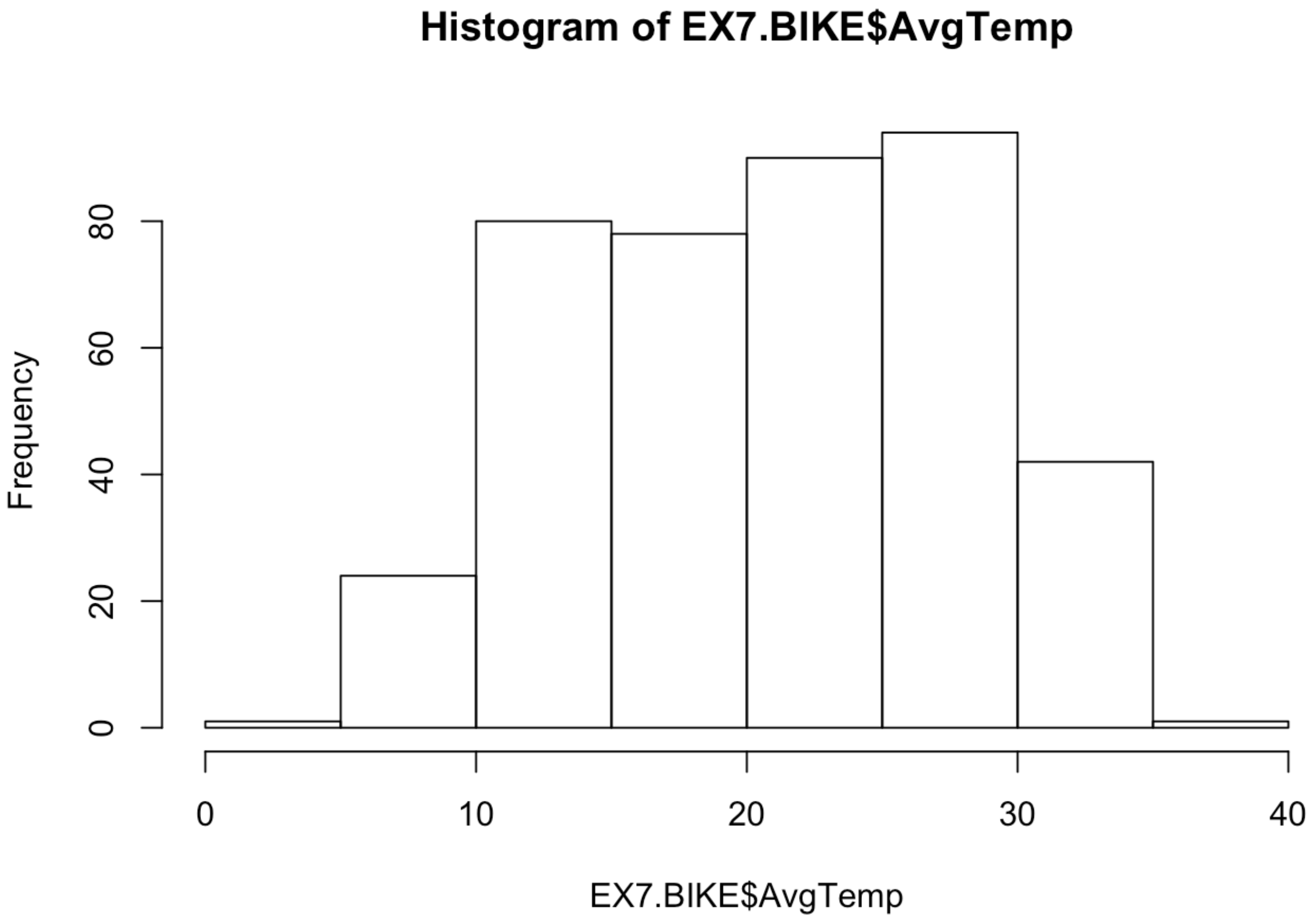
- At which of the four temperatures will the standard error of the predicted value be smallest? Explain. Hint: look at a `summary` of the `AvgTemp` column.

Comment: The standard error will be the smallest at temperature 20, as this is nearest the mean AvgTemp. ()

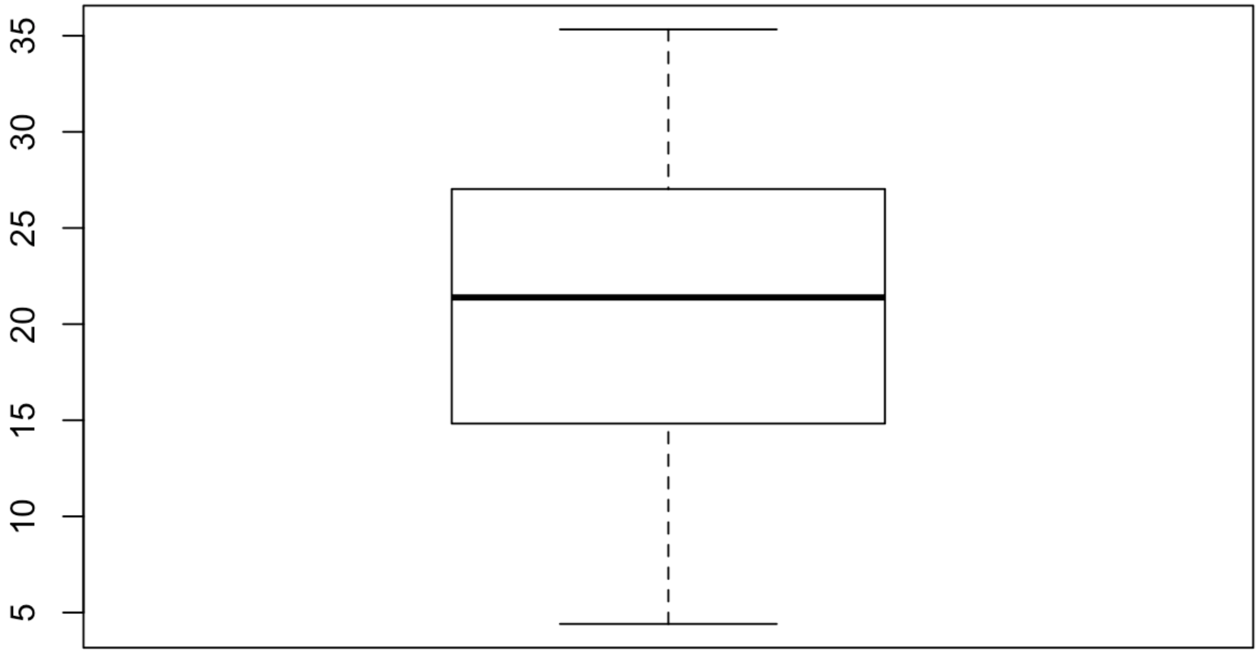
```
#code looking at summary of AvgTemp column
summary(EX7.BIKE$AvgTemp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   4.408  14.840  21.390  20.930  27.010  35.330
```

```
hist(EX7.BIKE$AvgTemp)
```



```
boxplot(EX7.BIKE$AvgTemp)
```



Question 2

Use the `data` command to read in the `ACCOUNT` dataset. Let us try to model the relationship between the amount of money someone has in their checking account `CheckingBalance` (x) and their savings account `SavingBalance` (y).

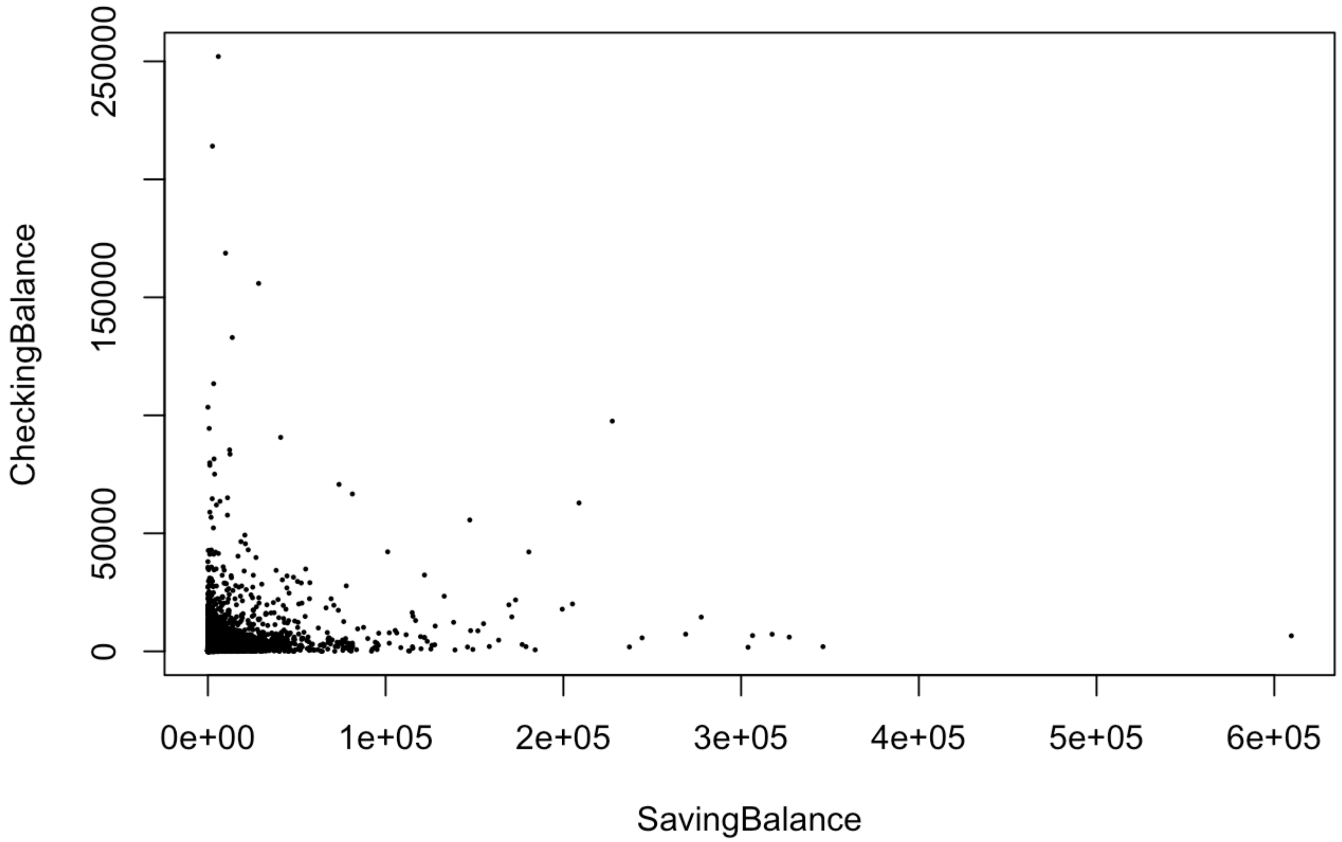
```
#code loading up the ACCOUNT dataset
data("ACCOUNT")
```

a. First, let us restrict ourselves to customers who have at some money in each account. Using `subset` , create a data frame called `AMOUNTS` which contains the rows of `ACCOUNT` that have both `CheckingBalance` and `SavingBalance` greater than 2.5. Hint: see slides 57-58 in the Rbasics.pdf lecture notes. Your `AMOUNTS` dataset ends up with exactly 9100 rows.

```
#code for taking subset of ACCOUNT and naming it AMOUNTS
AMOUNTS <- subset(ACCOUNT,CheckingBalance>2.5&SavingBalance>2.5)
```

b. Provide a scatterplot of the relationship from your newly created `AMOUNTS` data frame. Without even fitting the model, how do you know that a linear regression cannot be used to describe this relationship?

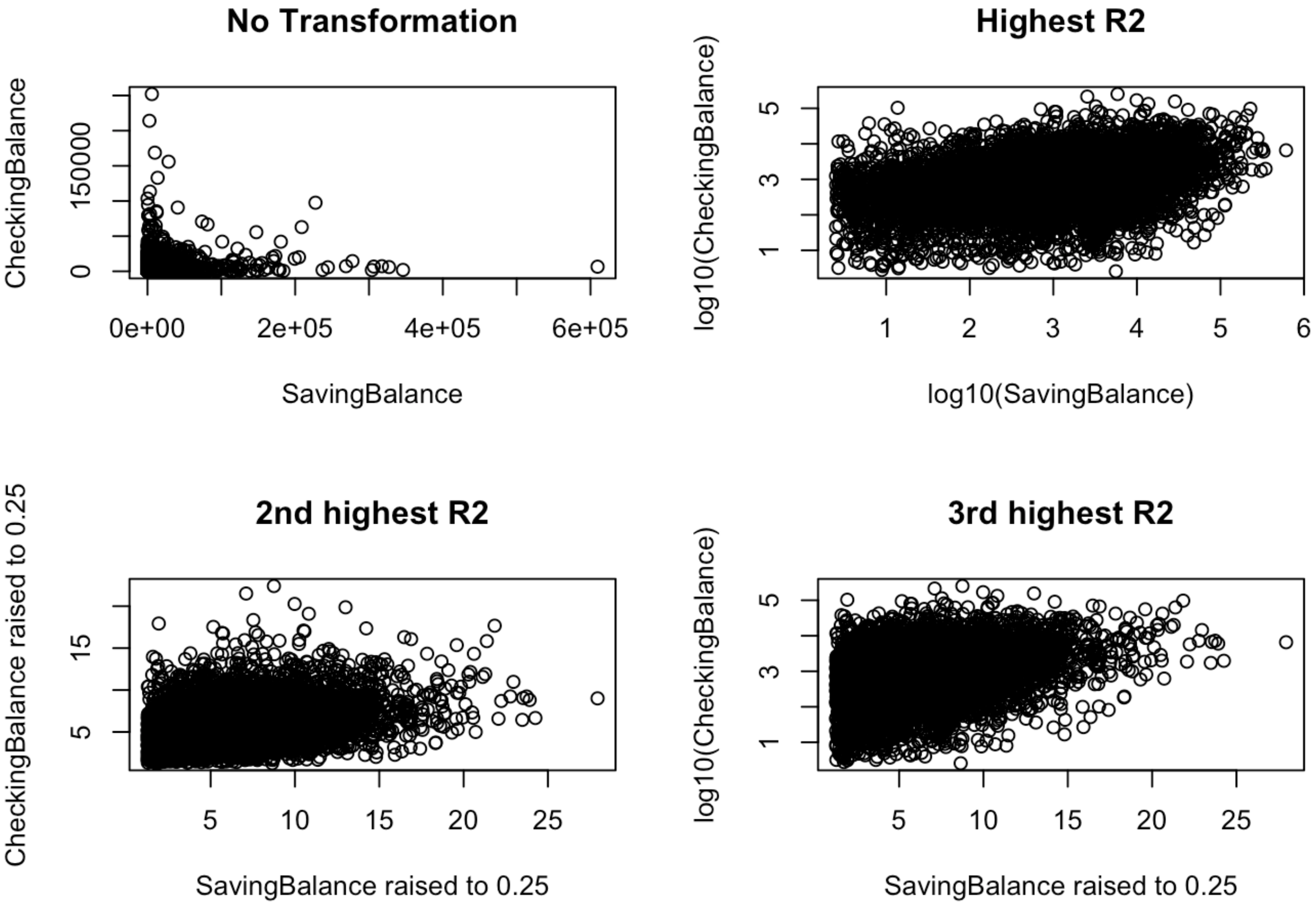
```
#code for scatterplot using AMOUNTS dataframe
plot(CheckingBalance~SavingBalance,data=AMOUNTS,pch=20,cex=0.3)
```



Comment: There is no relationship between a checking and a savings account. Many people do not have both. ()

c. Fit the model then run `find.transformations` (add an additional argument `pch=20` and `cex=0.3` to make the points small). Report the R^2 of the model using the original variables, the R^2 of the “best” transformation, and what set of transformations yields the model that best fits the data.

```
#code fitting the model on the AMOUNTS dataframe
N <- lm(CheckingBalance~SavingBalance,data=AMOUNTS)

#code running find.transformations
find.transformations(N)
## y is CheckingBalance
## x is SavingBalance
```



```
## No transformation yields rsquared of 0.027
##
##  x.power y.power rsquared
##    log10   log10    0.126
##     0.25    0.25    0.122
##     0.25   log10     0.12
##    log10    0.25    0.116
##    log10   -0.25    0.108
##    -0.25   log10    0.107
##      0.5    0.25    0.106
summary(N)
##
## Call:
## lm(formula = CheckingBalance ~ SavingBalance, data = AMOUNTS)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -35185  -1946  -1384   -165 249522
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.149e+03  7.699e+01   27.91   <2e-16 ***
## SavingBalance 6.499e-02  4.093e-03   15.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6965 on 9098 degrees of freedom
## Multiple R-squared:  0.02697,    Adjusted R-squared:  0.02686
## F-statistic: 252.1 on 1 and 9098 DF,  p-value: < 2.2e-16
```

Comment: Without any transformation, the rsquared value is 0.027. Highest R^2 with transformation is 0.126 with an x power and y power of log10, but even that is not very high.

d. Define `check.trans` and `save.trans` to be the transformed values of the original variables in the `AMOUNTS` dataset. Fit this model and report the "intercept" "slope" to one decimal place (e.g., 2.5, 0.9, etc.). Report the "untransformed" equation that written in the form `Savings = ...`

```
#code defining check.trans
check.trans <- log10(AMOUNTS$CheckingBalance)
#code defining save.trans
save.trans <- log10(AMOUNTS$SavingBalance)
#fitting the transformed model, i.e. M.trans <- lm(y.trans~x.trans)
M.trans <- lm(save.trans~check.trans)
#code getting the summary of the transformed model
summary(M.trans)
##
## Call:
## lm(formula = save.trans ~ check.trans)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1573 -0.5619  0.1093  0.6488  2.5904
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.47685    0.04329   34.11   <2e-16 ***
## check.trans  0.51847    0.01433   36.18   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9024 on 9098 degrees of freedom
## Multiple R-squared:  0.1258, Adjusted R-squared:  0.1257
## F-statistic:  1309 on 1 and 9098 DF,  p-value: < 2.2e-16
```

Intercept and slope of transformed regression: Savings = 0.07274*Checking* + *1.299*e^10

Untransformed Equation: Savings = .065*Checking + 2149

Question 4

On a separate piece of paper (unless you want to figure out how to write equations in this format), rewrite each equation so that it is in the form y = … Note: you do not need to simplify except for (c), which should be written as $y = ax^b$ for the correct values of a and b. See the Untransformation Primer.

a. $\sqrt{y} = 3 + 4/x\ y = (3 + 4/x)^2$

b. $\log_{10} y = 1 - 2x^2\ y = e^{(}1 - 2x^2)$

c. $\log_{10} y = 2 + 3\log_{10} x\ y = 3e^{(}2x)$

d. $1/y^2 = 1 - 2x^2\ y = 1/sqrt(1 - 2x^2)$

e. $y^2 = -4 + 8\sqrt{x}\ y = sqrt(-4 + 8sqrtx)$