Homework 4

46-921, Fall 2017

Due Thursday, September 28 at 3:00 PM

You should submit the Rmd file with your answers in the appropriate spaces. Rename the file as `YOURANDREWID_HW4.Rmd` and submit it via Canvas. Also submit the `.pdf` file that is produced.

Whenever you are asked to create a plot, unless the specific form is stated, you have the flexibility to use your judgement to choose the plot you feel is most appropriate. I expect that you will take steps to make the plot clean and readable.

I am assuming at this point you know how to appropriately specify R commands within this Markdown file, i.e., using the "triple quotation marks."

This exercise is based on data that appeared in the November 9, 1988 edition of the *Wall Street Journal*. The example was originally used in Siegel (1997), but was also used in Sheather (2009). To quote Siegel:

> US Treasury bonds are among the least risky investments, in terms of the likelihood of your receiving the promised payments. In addition to the primary market auctions by the Treasury, there is an active secondary market in which all outstanding issues can be traded. You would expect to see an increasing relationship between the coupon of the bond, which indicates the size of its periodic payment (twice a year), and the current selling price. The … data set of coupons and bid prices [are] for US Treasury bonds maturing between 1994 and 1998 … The bid prices are listed per 'face value' of $100 to be paid at maturity. Half of the coupon rate is paid every six months. For example, the first one listed pays $3.50 (half of the 7% coupon rate) every six months until maturity, at which time it pays an additional $100.

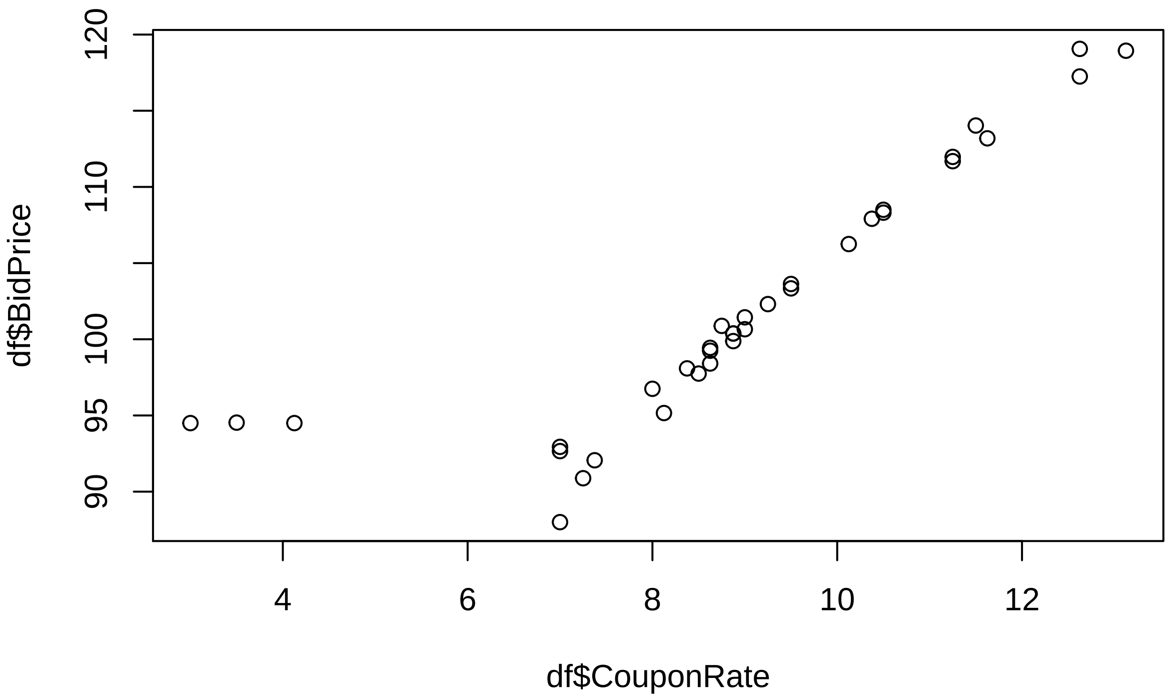
The data file can be found at http://www.stat.cmu.edu/~cschafer/MSCF/bonds.txt.

Do each of the following. We will treat "Bid Price" as the response and "Coupon Rate" as the predictor.

1. Create a scatter plot of the response versus the coupon rate. Comment on the form.

Solution

```r
df = read.table('bonds.txt', header=T)
plot(df$CouponRate,df$BidPrice)
```



Comments

- The response `BidPrice` and predictor `CouponRate` are positively correlated.
- There are three evident outliers that deviate from the linear pattern very much.

---

2. Fit a simple linear regression model to the data. Show the output from R, including the table of parameters and their standard errors. This information is found in R using the function `lm()`. The command `summary(holdout)` shows all of the relevant information.

Solution

```
model.1 = lm(data=df, BidPrice~ CouponRate)
summary(model.1)
```

```
##
## Call:
## lm(formula = BidPrice ~ CouponRate, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.249 -2.470 -0.838  2.550 10.515
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  74.7866     2.8267  26.458  < 2e-16 ***
## CouponRate    3.0661     0.3068   9.994 1.64e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 4.175 on 33 degrees of freedom
## Multiple R-squared:  0.7516, Adjusted R-squared:  0.7441
## F-statistic: 99.87 on 1 and 33 DF,  p-value: 1.645e-11
```

The model estimation is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_0 = 74.7866; \quad \hat{\beta}_1 = 3.0661$$

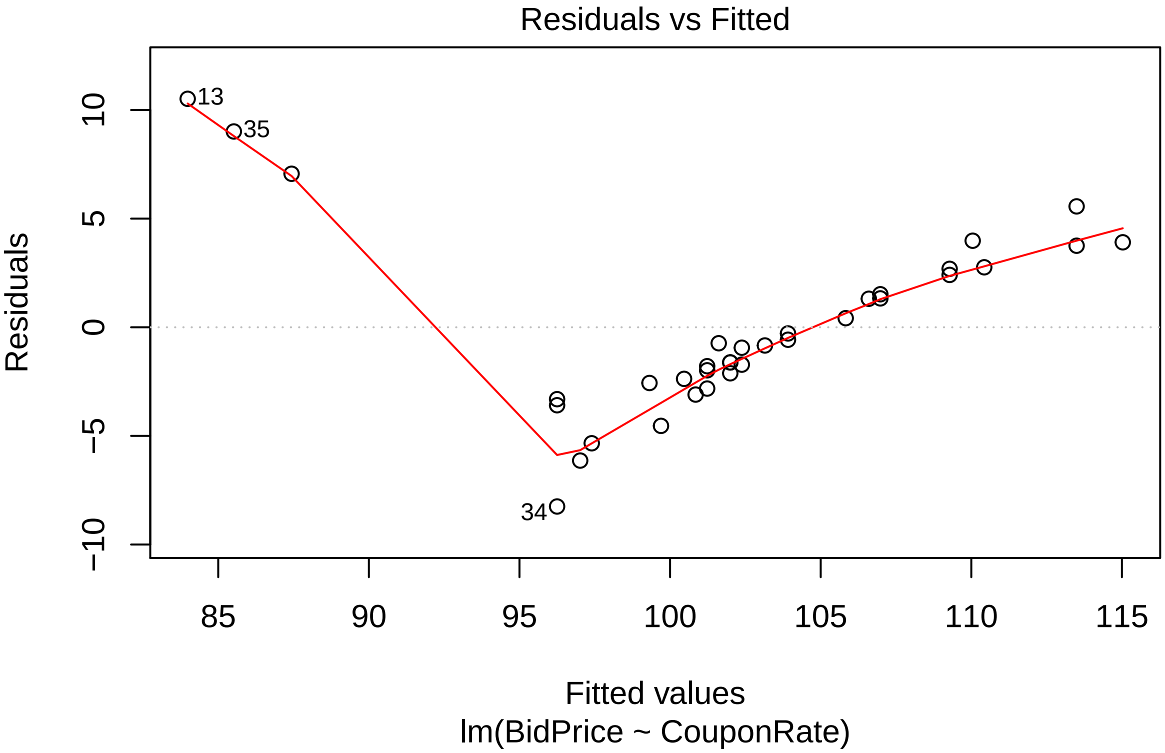
$$SE(\hat{\beta}_0) = 2.8267; \quad SE(\hat{\beta}_1) = 0.3068$$

---

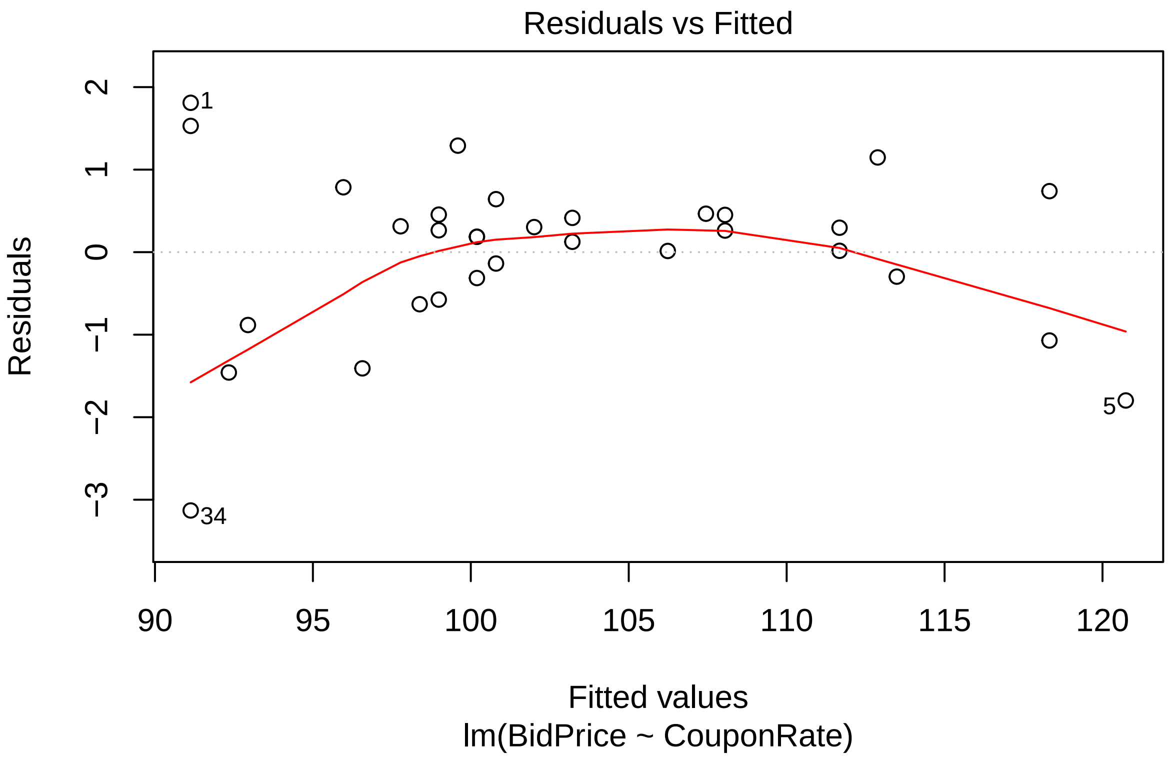
3. Construct the plot of residuals versus fitted values and comment on its form. The residuals can be obtained from `holdout$residuals` and the fitted values from `holdout$fitted.values`.

Solution

```r
# residual plot of the original model
plot(model.1, which=1)
```



```r
# residual plot of the model with outlier removed
df.trunc = df[df$CouponRate>5,]
model.2 = lm(data = df.trunc, BidPrice~CouponRate)
plot(model.2, which=1)
```



Comments

- The residual v.s. fitted value plot of the original model is clearly problematic because of the outliers.
- If we remove the three outliers and fit another model, residual v.s. fitted value plot is a still little curved, which implies that the relationship between `BidPrice` and `CouponRate` may be non-linear. But it's tricky to judge, we need more data to verify the existance (or non-existance) of curvature.

---

4. Construct a 95% confidence interval for β_1 based on this initial model. Do you trust this confidence interval?

Solution

```
# for the original model
confint(model.1)
```

```
##                  2.5 %    97.5 %
## (Intercept) 69.035683 80.537437
## CouponRate   2.441906  3.690299
```

So the 95% confidence intervals are

$$CI(\beta_0) = [9.035683, \quad 80.537437]$$

$$CI(\beta_1) = [2.441906, \quad 3.690299]$$

I do not trust this confidence interval for the original model, because the residual v.s. fitted value diagnostic plot is not "nice". Assume the true data generating process is

$$y = \beta_0 + x\beta_1 + u$$

Under $H_0 : \beta_1 = 0$, the quantity $T(\hat{\beta}_1)$ is calcuated as

$$T(\hat{\beta}_1) = \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1\sqrt{n-1}s_x}{\hat{\sigma}}$$

where $s_x^2 = \frac{1}{n-1}\tilde{\mathbf{x}}^\top\tilde{\mathbf{x}}$, $\tilde{\mathbf{x}} = \mathbf{x} - \bar{\mathbf{x}}$, and $\hat{\sigma} = RSS/(n-2)$. It follows that

$$T(\hat{\beta}_1) = \frac{\tilde{\mathbf{x}}^\top\tilde{\mathbf{y}}}{\sigma(\tilde{\mathbf{x}}^\top\tilde{\mathbf{x}})^{1/2}} \cdot \frac{1}{RSS/\sigma(n-2)} =: z_{\beta_1} \cdot \frac{1}{RSS/\sigma(n-2)}$$

Where σ^2 is the true variance of u, we mannully added it to normalize terms, σ cancels out in the formula above.

We can observe that $T(\hat{\beta}_1) \sim \textit{student.t}(n-2)$ **if and only if** $u \sim N(0, \sigma^2)$; in this case $z_{\beta_1} \sim N(0,1)$ and $RSS/\sigma(n-2) \sim \chi^2(n-2)$.

The residual v.s. fitted value plot suggests that u is not normally distributed. So the distributional assumption for hypothesis test does not hold. T quantity has little meaning.

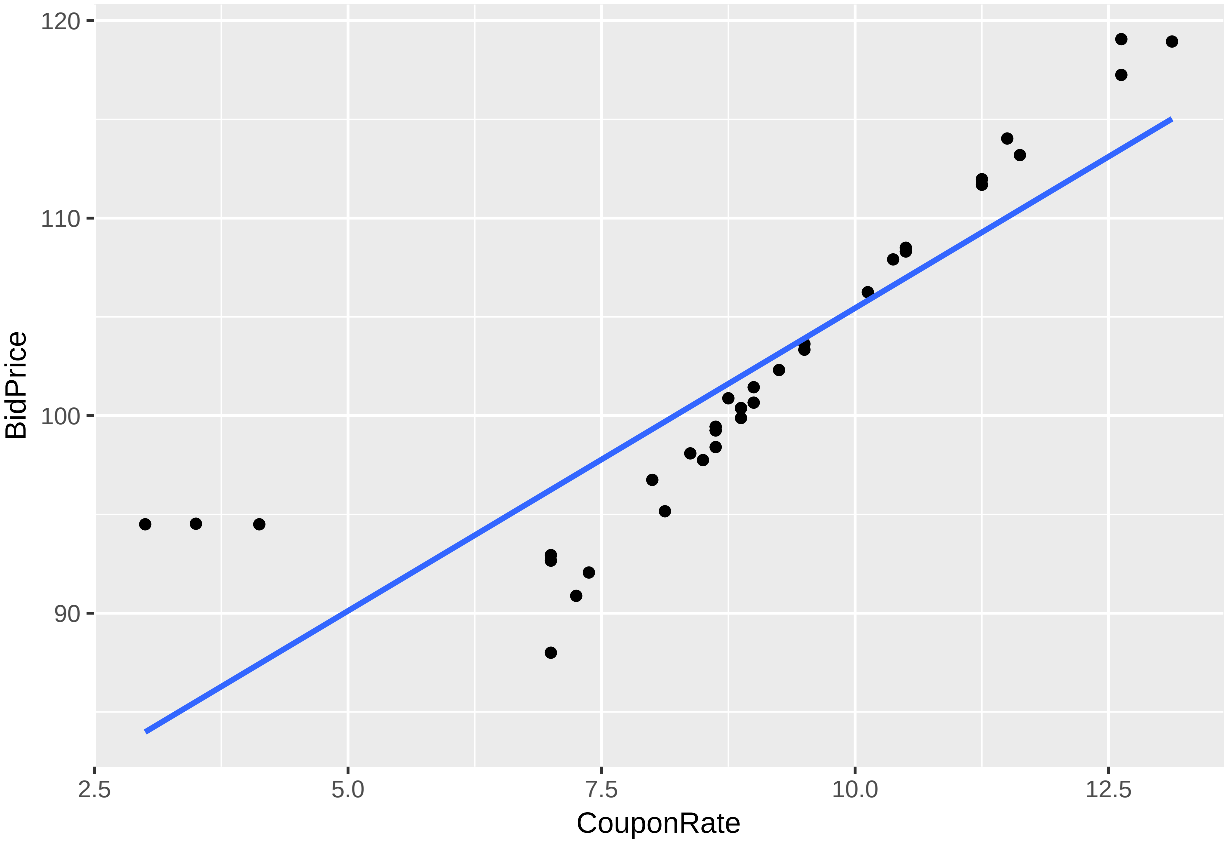
The trustworthiness of hypothesis test for modified model (with outliers removed) is somewhat tricky. The residual plot has a little curvature, but not obvious. We need more data to verify.

---

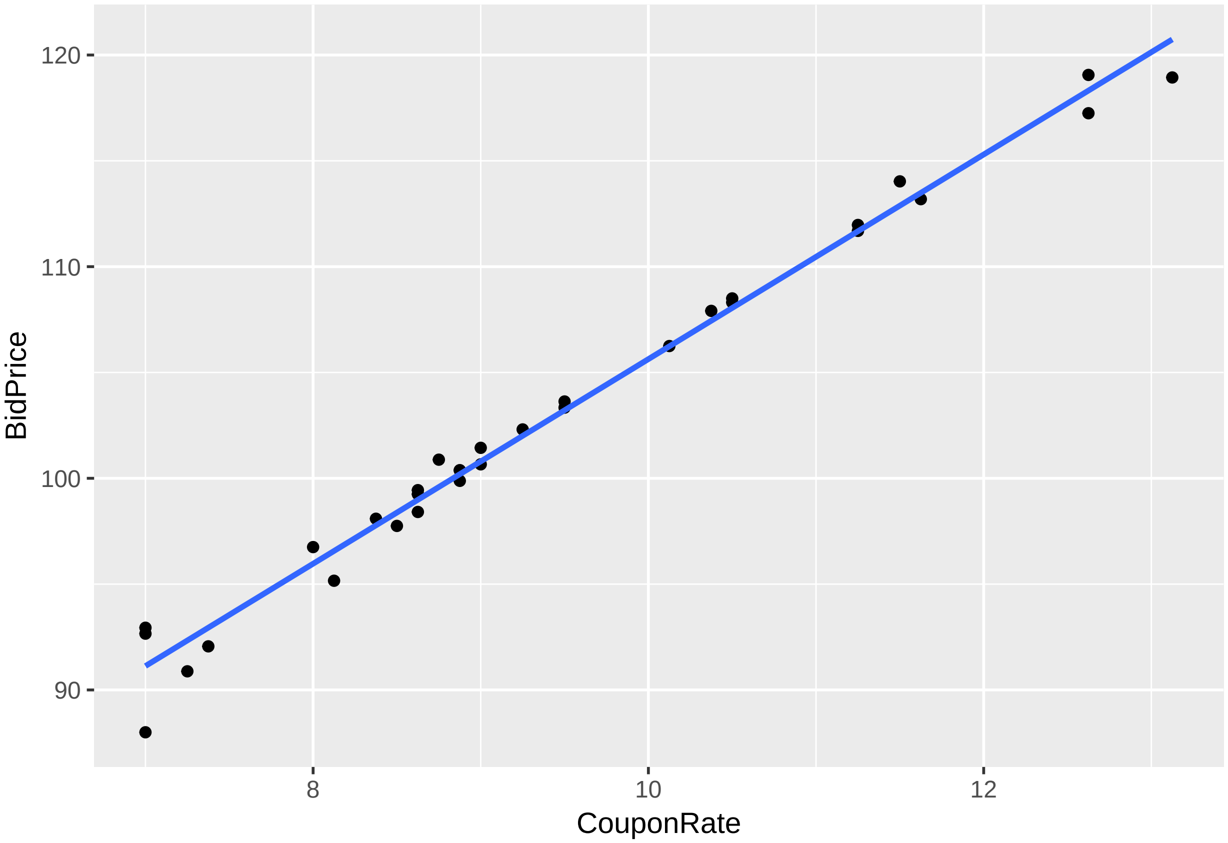
5. Look at the `ggplot2` function `geom_smooth()` and use it to create the scatter plot with the least squares regression line superimposed. (Hint: Use with `method="lm"`.)

Solution:

```r
# for the original model
ggplot(df, aes(CouponRate, BidPrice)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



```
# with outliers removed
ggplot(df.trunc, aes(CouponRate, BidPrice)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```


120
110
100
90
BidPrice
8
10
12
CouponRate