Simple Linear Regression
Homework Problems
Homework Solutions

Rob McCulloch

1.1. Problem: SLR Model

Suppose we are modeling house price as depending on house size. Price is measured in thousands of dollars and size is measured in thousands of square feet.
Suppose our model is:

$$P = 20 + 50s + \epsilon, \epsilon \sim N(0, 15^2).$$

That is, suppose that somehow we *know* the parameters: $\beta_0 = 20$, $\beta_1 = 50$, and $\sigma = 15$.
(a)
Given you know that a house has size s $= 1.6$, give a 95% predictive interval for the price of the house.
(b)
Given you know that a house has size s $= 2.2$, give a 95% predictive interval for the price.

(c)
In our model the slope is 50. What are the units of this number?
(d)
What are the units of the intercept 20?
(e)
What are the units of the standard deviation 15?
(f)
Suppose we change the units of price to dollars and size to square feet. What would the values and units of the intercept, slope, and error standard deviation?
(g)
If we plug s = 1.6 into our model equation (with the original units), P is a constant plus the normal random variable ϵ.

Given s = 1.6, what is the distribution of P ?

Solution

(a)
Given you know that a house has size $s = 1.6$, give a 95% predictive interval for the price of the house.
The point prediction is $P_f = 20 + 50 * 1.6 = 100$ (P_f is the "future P").
The prediction interval is $(100 \pm 2 * 15) = (70, 130)$.
(b)
Given you know that a house has size $s = 2.2$, give a 95% predictive interval for the price.
The point prediction is $P_f = 20 + 50 * 2.2 = 130$ The prediction interval is $[130 \pm 2 * 15] = [100, 160]$.
(c)
In our model the slope is 50. What are the units of this number?
\$1,000 / 1,000 Sq. Feet = \$/Sq. Feet

$1,000 (same as P)
(e)
What are the units of the the error standard deviation 15? $1,000 (same as P)
(f)
Suppose we change the units of price to dollars and size to square feet What would the values and units of the intercept, slope, and error standard deviation? Intercept: 20,000 $ Slope: 50 $/Sq. Feet error standard deviation: 15,000 $
(g)
If we plug $s = 1.6$ into our model equation, P is a constant plus the normal random variables . Given $s = 1.6$, what is the distribution of P ?
When $s = 1.6$ the mean of house prices is $20 + 50 * 1.6 = 100$. The error standard deviation is the same, 15. Therefore

$$P \mid S = 1.6 \sim \mathrm{N}(100, 15^2)$$

3.1. Problem: The Shock Absorber Data

The data comes from a company which supplies a major automobile manufacturer with shock absorbers. An important characteristic is the "force transferred through the shock absorber when the shank is forced out of the cylinder". If you don't know what that really means, don't worry, neither do I.

What we do need to understand is that the manufacturer only considers the shock to be an acceptable part if the force measurement is between 485 and 585.

The shock manufacturer and the auto manufacturer are arguing over the following issue. Before the shock is finally shipped, it is filled with gas. After it is filled with gas, it becomes very difficult to measure the force characteristic we are interested in. The shock manufacturers would like to make the measurement before the shock is filled with gas. The auto maker is concerned that there may be a difference in the force before and after the shock is filled with gas and so would like to make the measurement after it is filled.

The shock maker claims that there is little difference between the before and after measurement so that the before measurement can be used.

To investigate this we have the before (column 1, reboundb) and the after (column 2, rebounda) measurements on 35 shocks (in shock.csv).

Get the shock data (shock.csv) from the webpage.

(a)
Plot reboundb vs. rebounda.
Does this look like the kind of data the simple linear regression model is designed to capture?

```
Excel:
Download the file shock.csv and double click the file icon to get into excel.
Click on a cell in the data.
/Insert/Charts/the one with a picture of a scatter plot.
Right click on each axis  and then choose "Format Axis"
to change the plot range.
Play around with other plot options!!!

R:
sdat = read.csv("http://www.rob-mcculloch.org/data/shock.csv")
plot(sdat)
```

(b)
Run the regression of y=rebounda on x=reboundb.
What is the estimate of the true slope?

```
Excel:
Download the file shock.csv and double click the file icon to get into excel.
(i) /Data/Data Analysis/Regression
(ii) put in y range (e.g. b1:b36) and x range (e.g. a1:a36)
(iii) click labels and then OK.

R:
sdat = read.csv("http://www.rob-mcculloch.org/data/shock.csv")
#put the results of the regression in the data structure sreg (a list)
sreg = lm(rebounda~reboundb,sdat)
#print a summary of sreg
print(summary(sreg))
```

(c)
Given reboundb $= 535$, give the plug-in predictive interval for rebounda.

(d)
Give the 95% confidence interval for β_1.

(e)
Give the 95% confidence interval for β_0.

(f)
Test the null hypothesis (level .05) that $\beta_0 = 0$.

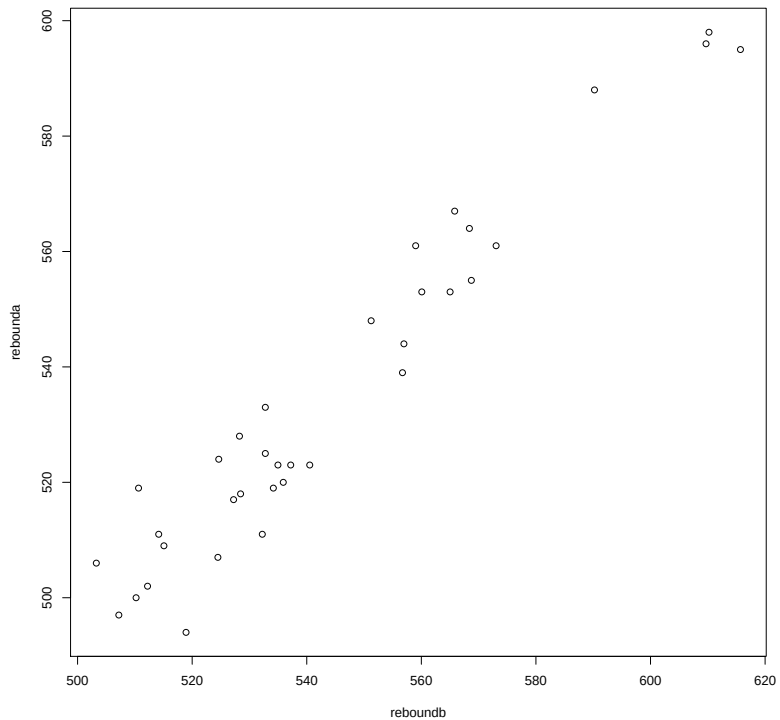
(g)
Test the null hypothesis (level .05) that $\beta_1 = 0$.

(h)
Test the null hypothesis (level .05) that $\beta_1 = 1$.
Why is this an interesting hypothesis to test?

(i)
Is it ok to use the before measurement as a proxy for the after measurement? What does the simple linear regression model tell us about this?

Solution

(a)



Yes, really looks like line + error !!

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.2259    23.8852   0.763    0.451
reboundb      0.9495     0.0438  21.675   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.67 on 33 degrees of freedom
Multiple R-squared:  0.9344,Adjusted R-squared:  0.9324
F-statistic: 469.8 on 1 and 33 DF,  p-value: < 2.2e-16
```

(b) $\hat{\beta}_1 = 0.9495$.

(c)

```
> 18.2259+0.9495*535+ 7.67*2*c(-1,1)
[1] 510.8684 541.5484
```

(d)

```
> 0.9495 + 2*0.0438*c(-1,1)
[1] 0.8619 1.0371
```

(e)

```
> 18.2259 + 2* 23.8852*c(-1,1)
[1] -29.5445  65.9963
```

(f)

pval = .451, fail to reject

(g)

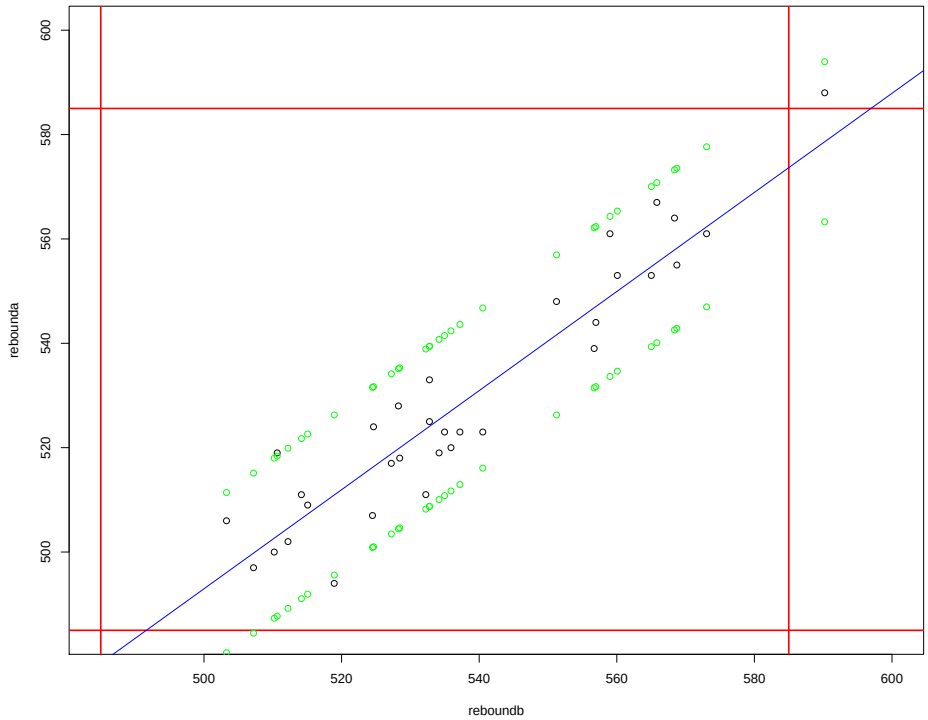
pval = 0, reject.

(h)

```
> t = ( 0.9495-1)/0.0438
> t
[1] -1.152968
```

fail to reject.
slope=1 and intercept=0 would be consistent with using the before as a proxy for the after.

(i)
Red is tolerance limits of (485,585).
Green is plug-in predictive intervals.
Blue is fitted line.



Short answer is yes. If before is in, you can be pretty sure after is in. Might want to double check if before is close to one of the limits.

3.2. Problem: Predictive Interval for the Shock Data

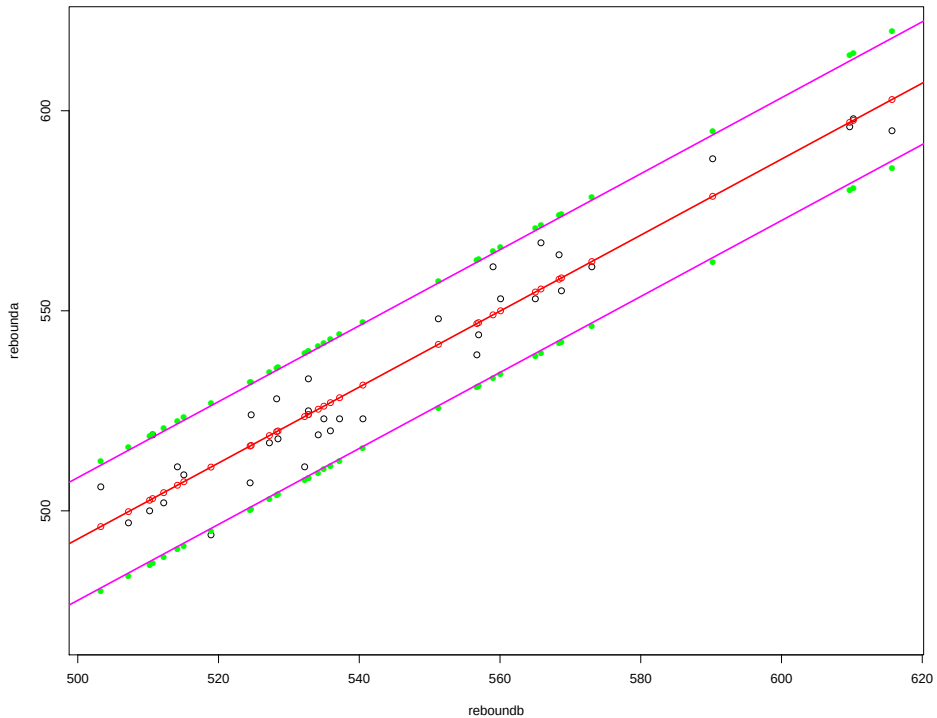
Let's compare the plug in predictive interval with the "correct" predictive interval (the one the accounts for our estimation error) for the shocks data.

Is there enough information in the data to make the plug-in interval similar to the predictive interval??

Here is the R code to get and plot the predictive and plug-in intervals.

```
sd = read.csv("shock.csv")
lms = lm(rebounda~reboundb,sd)

#note: try > ?predict.lm

predint = predict(lms,sd,interval="prediction")

plot(sd,ylim=c(470,620))
points(sd$reboundb,predint[,"fit"],col="red") #note predint is a matrix, not a data.frame
points(sd$reboundb,predint[,"lwr"],col="green",pch=16)
points(sd$reboundb,predint[,"upr"],col="green",pch=16)

sigmahat = summary(lms)$sigma #estimate of sigma
ab = coef(lms) #estimates of intercept and slope
abline(ab[1]+2*sigmahat,ab[2],col="magenta",lwd=2)
abline(ab[1]-2*sigmahat,ab[2],col="magenta",lwd=2)
abline(ab[1],ab[2],col="red",lwd=2)
```

Here is the plot.
Magenta is plug-in, green dots are predictive.



Is using the simple plug-in predictive a reasonable approach in this problem?

Solution

Yes.

3.3. Problem: Beta for Fidelity Funds

Get the data in the file fidrets.csv.

The data is monthly returns on: sp500: the s&p 500.
FidInc: a Fidelity income fund.
FidVal: a Fidelity "value" fund.
FidTech: a Fidelity Tech fund.
From the names, we might expect the value fund to be riskier than the income fund and the tech fund to be riskier than the value fund.

(a)

For each of the three funds plot sp500 vs. fund return.
Is linear regression a good way to think about the relationship between the market returns and the fund returns?

So that you can compare the different funds, make sure each plot is on the same scale.

(b)

For each of the three funds compute the 95% confidence interval for the slope.

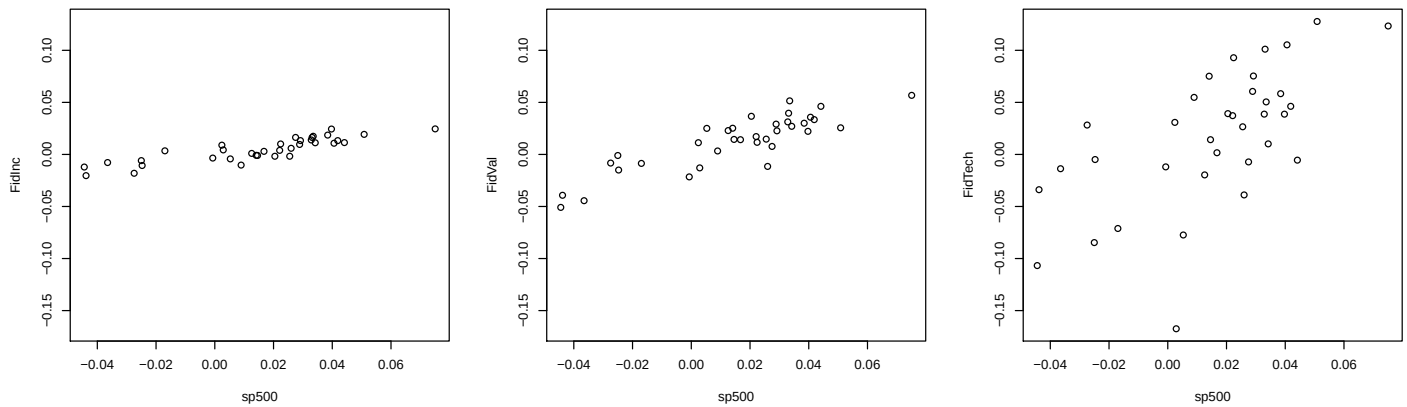
What do these intevals test us about the risk of the funds?

(c)

This is not well motivated in the example, but just as an excercise for each of the three funds test $\beta_1 = 0$ and $\beta_1 = 0$ (the slope is 0 and the slope is 1).

Solution

(a)



Wow, there is a huge difference and it certainly looks like the risk goes the way we anticipated.

In each case a linear regression seems reasonable.

Income regression.

```
Call:
lm(formula = FidInc ~ sp500, data = fr)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0127402 -0.0034358 -0.0009288  0.0041676  0.0109666

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0006729  0.0011224  -0.599    0.553
sp500        0.3550981  0.0356225   9.968 1.75e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.005842 on 33 degrees of freedom
Multiple R-squared:  0.7507,Adjusted R-squared:  0.7431
F-statistic: 99.37 on 1 and 33 DF,  p-value: 1.753e-11
```

Value regression.

```
[1] 0.2838531 0.4263431
al ~ sp500, data = fr)

Residuals:
      Min        1Q    Median        3Q       Max
-0.033008 -0.006692 -0.000823  0.009432  0.023928

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0005589  0.0025190   0.222    0.826
sp500       0.8052956  0.0799444  10.073 1.35e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.01311 on 33 degrees of freedom
Multiple R-squared:  0.7546,Adjusted R-squared:  0.7472
F-statistic: 101.5 on 1 and 33 DF,  p-value: 1.348e-11
```

Tech regression.

```
Call:
lm(formula = FidTech ~ sp500, data = fr)

Residuals:
      Min        1Q    Median        3Q       Max
-0.166180 -0.033587  0.005578  0.037878  0.075195

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.005604   0.009821  -0.571    0.572
sp500        1.506359   0.311696   4.833 3.02e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.05112 on 33 degrees of freedom
Multiple R-squared:  0.4144,Adjusted R-squared:  0.3967
F-statistic: 23.36 on 1 and 33 DF,  p-value: 3.017e-05
```

(b)

```
> #Income
> 0.3550981+ 2*0.0356225*c(-1,1)
[1] 0.2838531 0.4263431
>
> #Value
> 0.8052956 + 2*0.0799444*c(-1,1)
[1] 0.6454068 0.9651844
>
> #Tech
> 1.506359 + 2*0.311696*c(-1,1)
[1] 0.882967 2.129751
```

Seems like there is pretty clear evidence that (as least as far as "beta" goes) the Income fund is less risky than the other two.

Results suggest that Tech is the most risky but the uncertainty in estimation of the Tech slope is huge.

(c)
For testing slope=0 all the p-values are tiny =¿ reject.

For testing =1:

```
> #Income t
> (0.3550981-1)/0.0356225
[1] -18.10378
> #reject
>
> #Value t
> (0.8052956-1)/0.0799444
[1] -2.435498
> #reject
>
> #Tech t
> (1.506359-1)/0.311696
[1] 1.624528
> #fail to reject at usual levels.
> 2*pnorm(-1.624528)
[1] 0.1042632
> #p-value is about .1
```

The R code:

```
fr = read.csv("fidrets.csv")

par(mfrow=c(1,3))
for(i in 2:4) {
   plot(fr[,1],fr[,i],xlab="sp500",ylab=names(fr)[i],ylim=range(fr))
}

lmI = lm(FidInc~sp500,fr)
lmV = lm(FidVal~sp500,fr)
lmT = lm(FidTech~sp500,fr)

print(summary(lmI))
print(summary(lmV))
print(summary(lmT))
```