Multiple regression example: Ride share forecasting

Peter Hoff

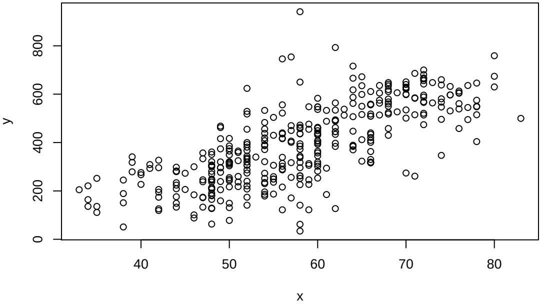
STAT 423

Applied Regression and Analysis of Variance

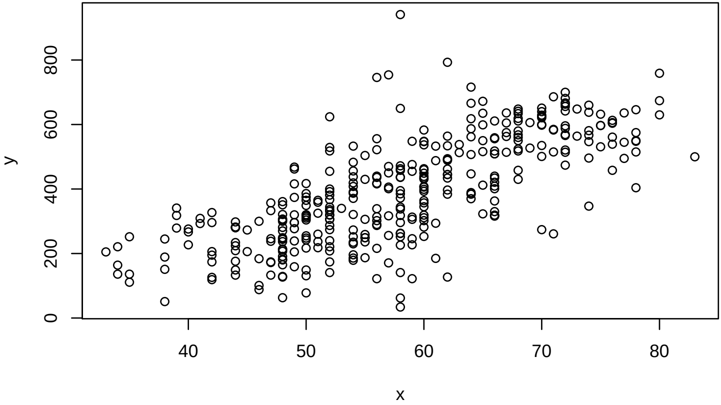
University of Washington

Forecast tomorrow's trip total:

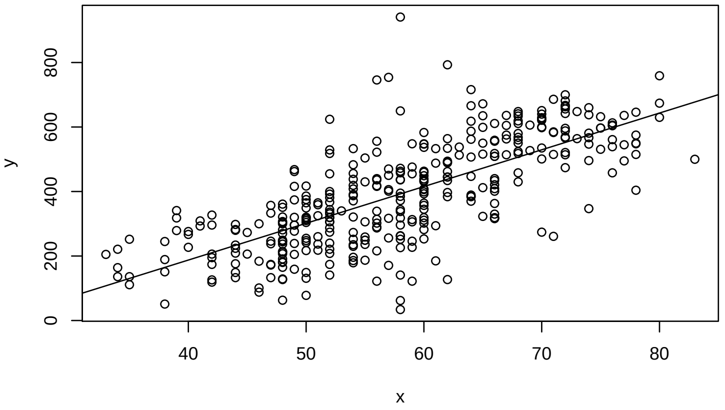
```
x<-weather$Mean_Temperature_F[-nrow(weather)]
y<-weather$ttrips[-1]

cor(x,y)

## [1] 0.7239651
```



```r
fit<-lm(y~x)

summary(fit)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -358.59  -74.64    8.42   71.70  548.41
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -268.4622    33.5743  -7.996 1.75e-14 ***
## x             11.3975     0.5708  19.968  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 113.6 on 362 degrees of freedom
## Multiple R-squared:  0.5241, Adjusted R-squared:  0.5228
## F-statistic: 398.7 on 1 and 362 DF,  p-value: < 2.2e-16
```



y
0
200
400
600
800
40
50
60
70
80
x



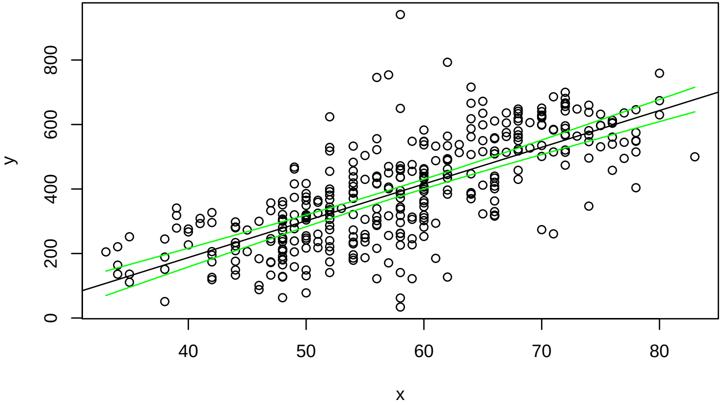
0
200
400
600
800
y
40
50
60
70
80
x

Prediction bands:

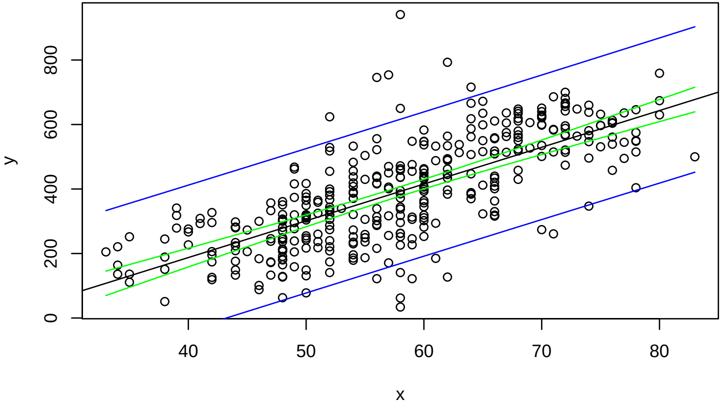
```
n<-length(x)
sigma<-sqrt( sum(fit$res^2)/(n-2) )
xbar<-mean(x)
SXX<-sum( (x-xbar)^2 )
```

```
xseq<-seq(min(x),max(x),by=1)

fit_x<-fit$coef[1] + fit$coef[2]*xseq

se_fit<- sigma*sqrt( 1/n + (xseq-xbar)^2/SXX )

se_prd<- sigma*sqrt( 1/n + (xseq-xbar)^2/SXX + 1 )
```

Q: Explain the difference between `se_fit` and `se_prd`.



0
200
400
600
800
y
40
50
60
70
80
x



0
200
400
600
800
y
40
50
60
70
80
x

Explained variation

Problem:

- $\hat{\sigma}^2$ is too big;
- $\hat{\sigma}^2$ measures variation in y not explained by x;
- maybe x isn't explaining much of the variation in y.

Explained variation

Problem:

- $\hat{\sigma}^2$ is too big;
- $\hat{\sigma}^2$ measures variation in y not explained by x;
- maybe x isn't explaining much of the variation in y.

$$SSY = \sum (y_i - \bar{y})^2 = \text{total variation in } y$$

$$RSS = \sum (y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_i])^2 = \text{variation in } y \text{ unexplained by } x$$

(**Exercise:** Show that $SSY \geq RSS$.)

Explained variation

Problem:

- $\hat{\sigma}^2$ is too big;
- $\hat{\sigma}^2$ measures variation in y not explained by x;
- maybe x isn't explaining much of the variation in y.

$$SSY = \sum (y_i - \bar{y})^2 = \text{total variation in } y$$
$$RSS = \sum (y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_i])^2 = \text{variation in } y \text{ unexplained by } x$$

(**Exercise:** Show that $SSY \geq RSS$.)

$$\begin{aligned} SSReg &= SSY - RSS = \text{variation in } y \text{ explained by } x \\ R^2 &= SSReg/SSY \\ &= (SSY - RSS)/SSY \\ &= 1 - RSS/SSY = \text{fraction of variation in } y \text{ explained by } x \end{aligned}$$

R^2 is often called the *coefficient of determination.*

```r
1- sum(fit$res^2)/sum( (y-mean(y))^2 )

## [1] 0.5241255

summary(fit)$r.squared

## [1] 0.5241255
```

Are there other variables that might help explain tomorrow's trip total?

```r
1- sum(fit$res^2)/sum( (y-mean(y))^2 )
```

```
## [1] 0.5241255
```

```r
summary(fit)$r.squared
```

```
## [1] 0.5241255
```

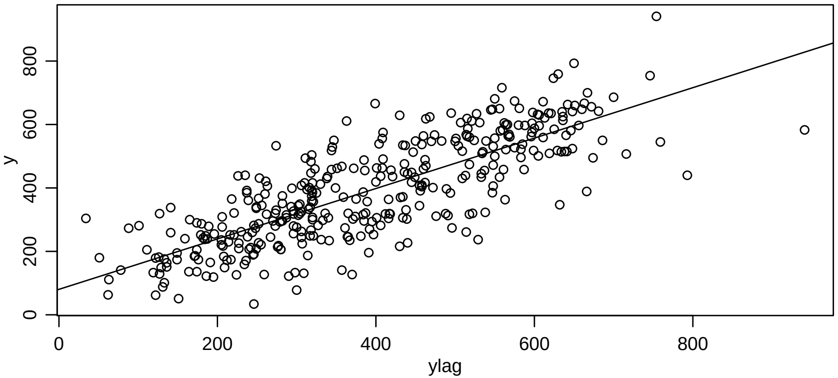
Are there other variables that might help explain tomorrow's trip total?

```r
cor(x,y)
```

```
## [1] 0.7239651
```

```r
ylag<-weather$ttrips[ -nrow(weather) ]
cor(ylag,y)
```

```
## [1] 0.7947798
```




```
fit1<-lm(y~x)

fit2<-lm(y~ylag)

summary(fit1)$r.squared

## [1] 0.5241255

summary(fit2)$r.squared

## [1] 0.6316749
```

Combining variables:

- x explains 52% of the variation in y
- *ylag* explains 63% of the variation in y

How much variation can be explained by both of them combined?

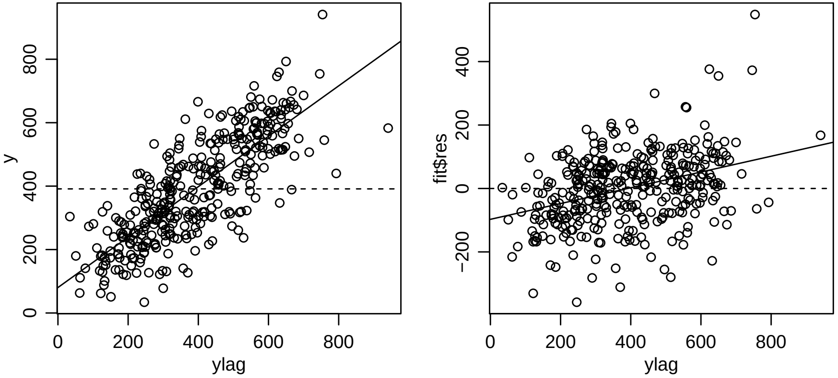
Combining variables:

- *x* explains 52% of the variation in *y*
- *ylag* explains 63% of the variation in *y*

How much variation can be explained by both of them combined?

```
fit3<-lm(y ~ x + ylag )

summary(fit3)$r.squared

## [1] 0.6676053
```

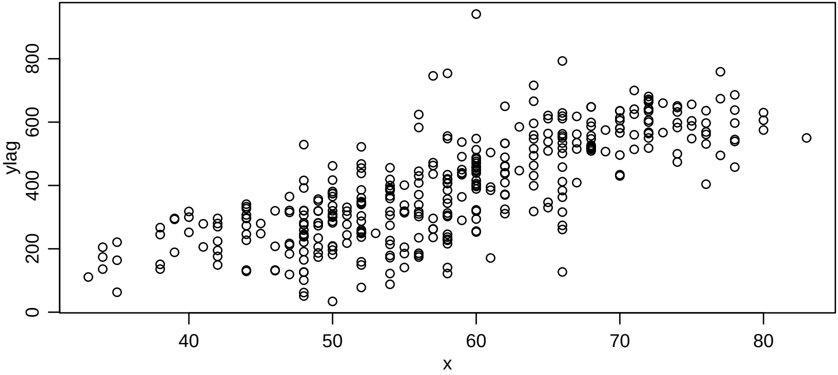
Added variable plots




```
cor(y, ylag)

## [1] 0.7947798

cor(y, fit1$res)

## [1] 0.6898366
```

Correlated predictors:

```
cor(x,ylag)

## [1] 0.7543267
```



Intuitively, if *ylag* and *x* were perfectly correlated, then the improvement in fit would be zero.

Prediction error improvement:

Recall in simple linear regression,

$$se(y^* - \hat{y}^*) = \hat{\sigma}\sqrt{1 + 1/n + (x^* - \bar{x})^2/SSX}.$$

If n is large, this will be dominated by $\hat{\sigma}$.

```
sqrt( sum(fit1$res^2)/(n-2) )

## [1] 113.5972

sqrt( sum(fit3$res^2)/(n-3) )

## [1] 95.07119

4*( sqrt( sum(fit1$res^2)/(n-2) ) - sqrt( sum(fit3$res^2)/(n-3) ) )

## [1] 74.10384
```

Adding more variables:

```
colnames(X)

##  [1] "Max_Temperature_F"            "Mean_Temperature_F"
##  [3] "Min_TemperatureF"             "Max_Dew_Point_F"
##  [5] "MeanDew_Point_F"              "Min_Dewpoint_F"
##  [7] "Max_Humidity"                 "Mean_Humidity"
##  [9] "Min_Humidity"                 "Max_Sea_Level_Pressure_In"
## [11] "Mean_Sea_Level_Pressure_In" "Min_Sea_Level_Pressure_In"
## [13] "Mean_Visibility_Miles"        "Min_Visibility_Miles"
## [15] "Max_Wind_Speed_MPH"           "Mean_Wind_Speed_MPH"
## [17] "Max_Gust_Speed_MPH"           "Precipitation_In"
## [19] "ylag"
```

Models we've tried:

```
Xa<-X[ ,c("Mean_Temperature_F") ]

summary(lm(y~Xa))$sigma

## [1] 113.5972

sqrt( mse_cv1(y,Xa) )

## [1] 113.8067
```

```
Xa<-X[ ,c("Mean_Temperature_F","ylag") ]

summary(lm(y~Xa))$sigma

## [1] 95.07119

sqrt( mse_cv1(y,Xa) )

## [1] 95.5772
```

All the variables:

```
summary(lm(y~X))$sigma

## [1] 90.21607

sqrt( mse_cv1(y,X) )

## [1] 93.08514
```

```
bigfit<-lm(y~X)

round( summary(bigfit)$coef , 3)

##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -3932.663   1044.015  -3.767    0.000
## XMax_Temperature_F              7.486      4.898   1.529    0.127
## XMean_Temperature_F           -15.552      9.185  -1.693    0.091
## XMin_TemperatureF               7.602      5.078   1.497    0.135
## XMax_Dew_Point_F               -2.554      3.426  -0.746    0.456
## XMeanDew_Point_F                4.643      6.160   0.754    0.452
## XMin_Dewpoint_F                 3.651      2.670   1.367    0.172
## XMax_Humidity                  -2.555      1.626  -1.572    0.117
## XMean_Humidity                 -0.380      2.789  -0.136    0.892
## XMin_Humidity                  -2.767      1.276  -2.169    0.031
## XMax_Sea_Level_Pressure_In    397.941    149.484   2.662    0.008
## XMean_Sea_Level_Pressure_In  -759.032    251.039  -3.024    0.003
## XMin_Sea_Level_Pressure_In    502.870    126.771   3.967    0.000
## XMean_Visibility_Miles          6.015      6.854   0.878    0.381
## XMin_Visibility_Miles          -6.295      2.906  -2.166    0.031
## XMax_Wind_Speed_MPH             1.358      2.778   0.489    0.625
## XMean_Wind_Speed_MPH            2.038      3.175   0.642    0.521
## XMax_Gust_Speed_MPH            -0.761      0.711  -1.070    0.285
## XPrecipitation_In              75.263     30.317   2.483    0.014
## Xylag                           0.502      0.061   8.176    0.000
```

Reduced model:

```
Xr<-X[ , summary(bigfit)$coef[-1,4] < .1 ]

summary(lm(y~Xr))$sigma

## [1] 91.24097

sqrt( mse_cv1(y,Xr) )

## [1] 92.52029
```