Collinearity

Peter Hoff

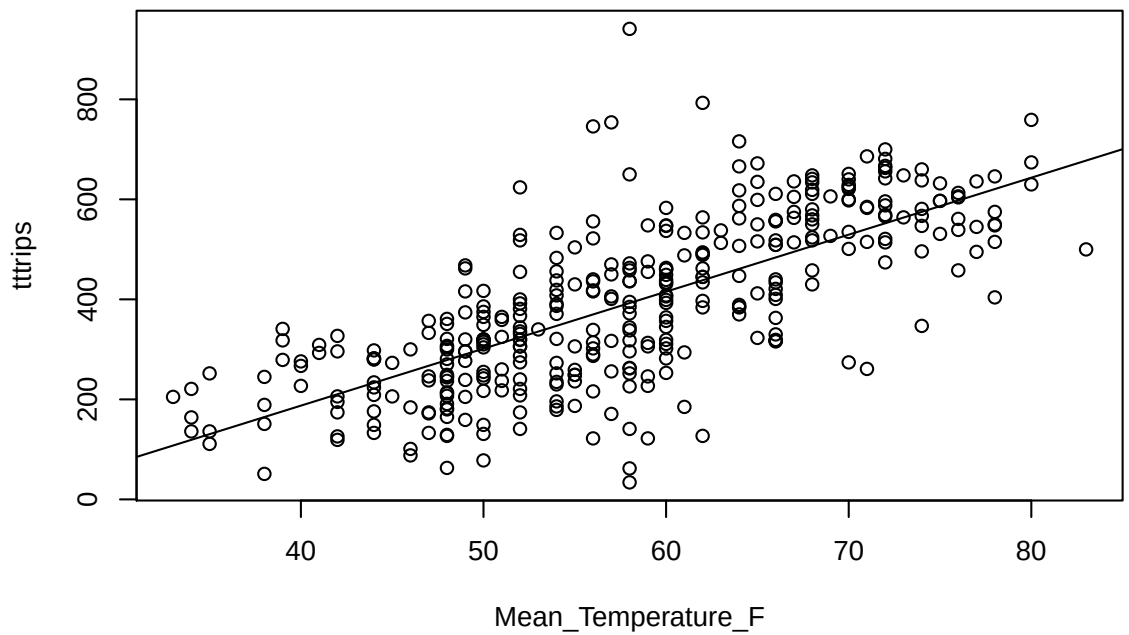
STAT 423

Applied Regression and Analysis of Variance

University of Washington

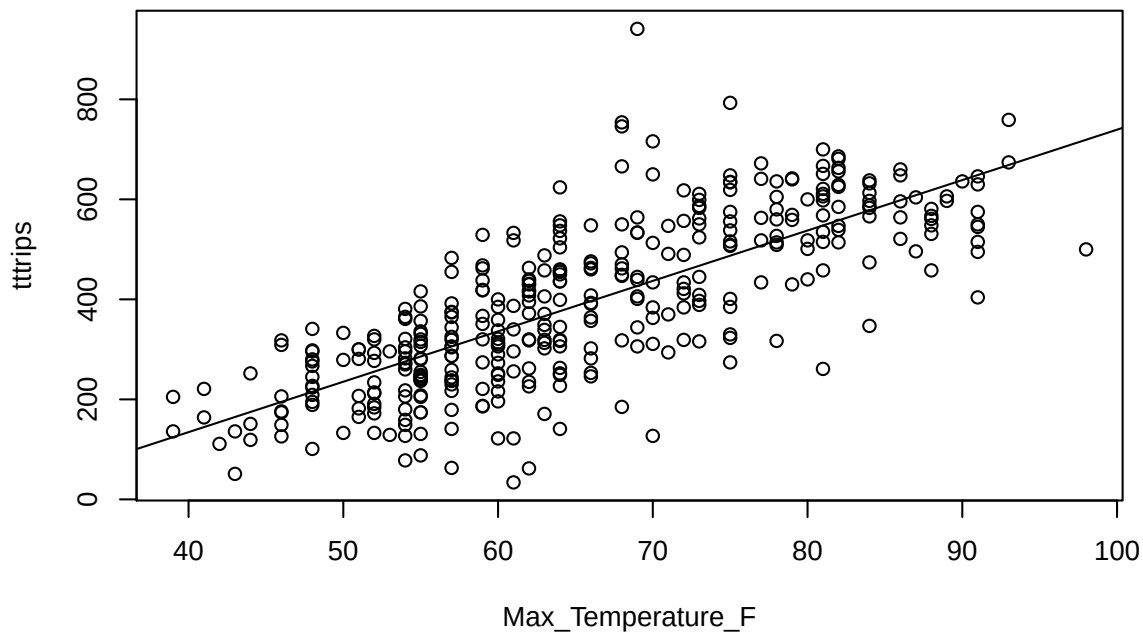
Forecast tomorrow's trip total:

```
summary( lm( tttrips ~ Mean_Temperature_F , data=weather ) )$coef

##                      Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)        -268.46222 33.5743459 -7.996052 1.747181e-14
## Mean_Temperature_F   11.39753  0.5708012 19.967592 2.441323e-60
```



Forecast tomorrow's trip total:

```
summary( lm( tttrips ~ Max_Temperature_F , data=weather ) )$coef

##                      Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)       -268.78746 30.4292681 -8.833189 4.460137e-17
## Max_Temperature_F   10.08121  0.4566869 22.074671 5.117409e-69
```



Forecast tomorrow's trip total:

```
summary( lm( tttrips ~ Mean_Temperature_F + Max_Temperature_F ,
             data=weather) )$coef

##                       Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)        -251.829061  31.773619 -7.925728 2.851907e-14
## Mean_Temperature_F   -4.274695   2.381626 -1.794864 7.351178e-02
## Max_Temperature_F    13.601363   2.013388  6.755459 5.717266e-11
```

Forecast tomorrow's trip total:

```
summary( lm( tttrips ~ Mean_Temperature_F + Max_Temperature_F ,
             data=weather) )$coef

##                       Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)        -251.829061  31.773619 -7.925728 2.851907e-14
## Mean_Temperature_F   -4.274695   2.381626 -1.794864 7.351178e-02
## Max_Temperature_F    13.601363   2.013388  6.755459 5.717266e-11
```

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- y = total trips tomorrow
- x_1 = today's mean temperature
- x_2 = today's max temperature

Why did the effect "change"?

Forecast tomorrow's trip total:

```
summary( lm( tttrips ~ Mean_Temperature_F + Max_Temperature_F ,
             data=weather) )$coef

##                       Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)        -251.829061  31.773619 -7.925728 2.851907e-14
## Mean_Temperature_F   -4.274695   2.381626 -1.794864 7.351178e-02
## Max_Temperature_F    13.601363   2.013388  6.755459 5.717266e-11
```

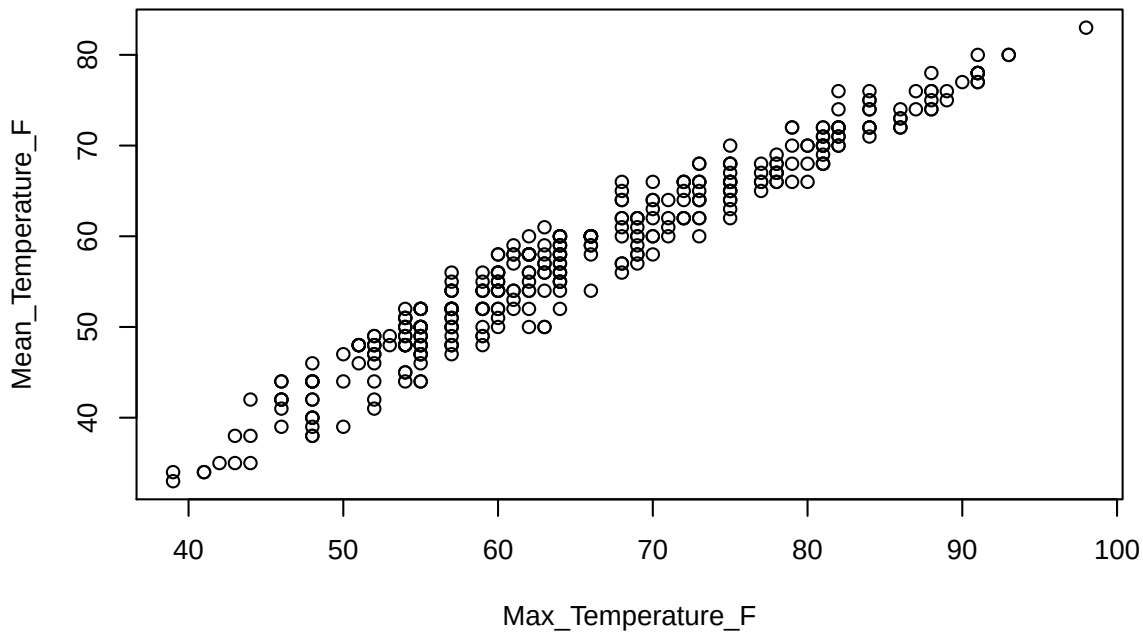
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- y = total trips tomorrow
- x_1 = today's mean temperature
- x_2 = today's max temperature

Why did the effect "change"?

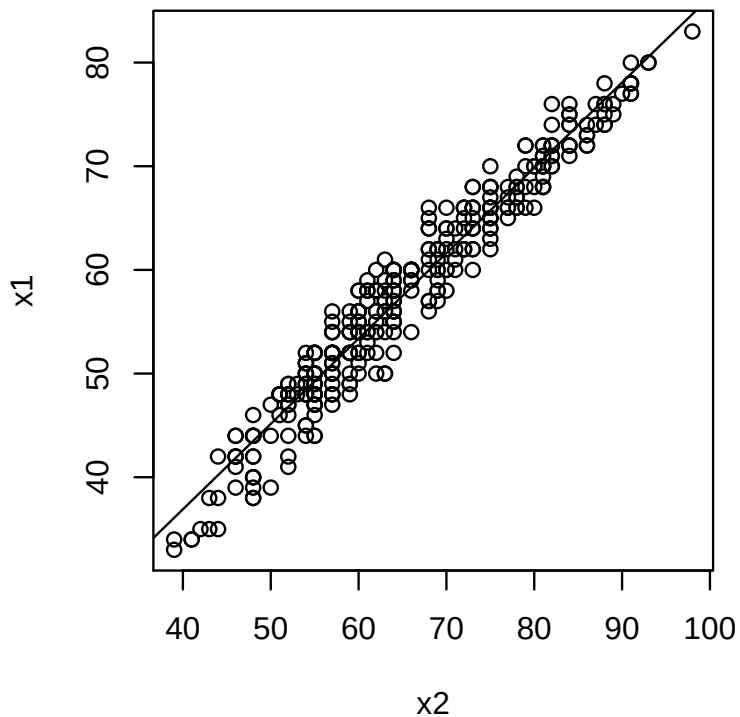
How do we interpret $\hat{\beta}_2$?

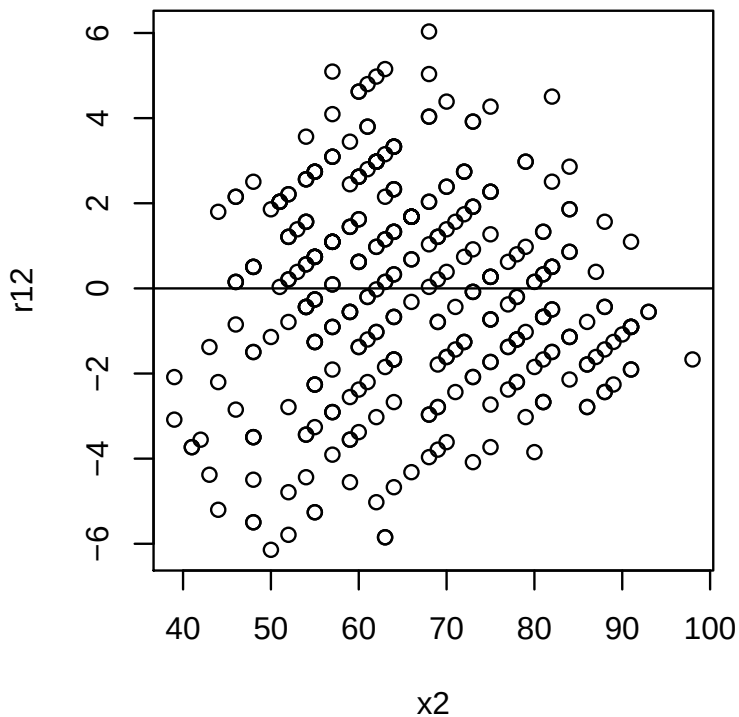
Collinearity



Collinearity

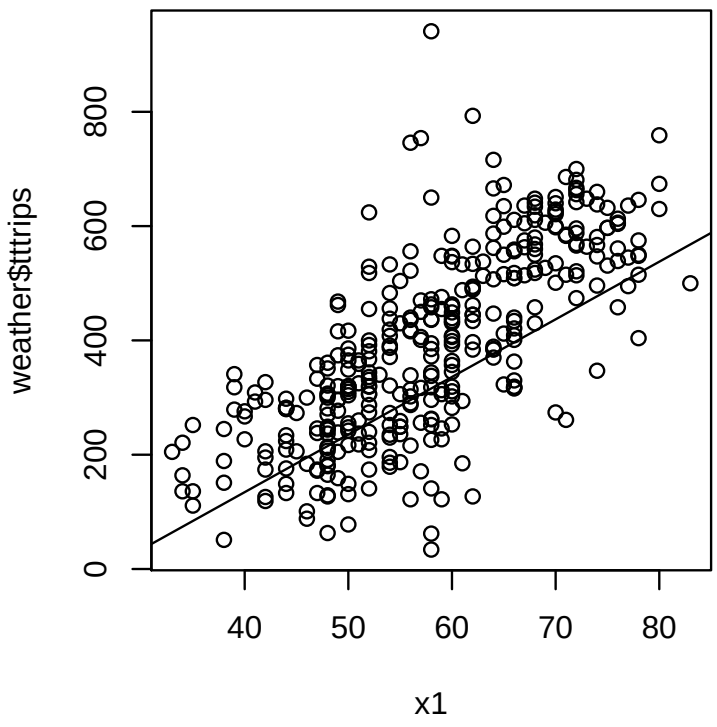
```
x1<-weather$Mean_Temperature_F
x2<-weather$Max_Temperature_F
r12<-lm( x1 ~ x2)$res
```

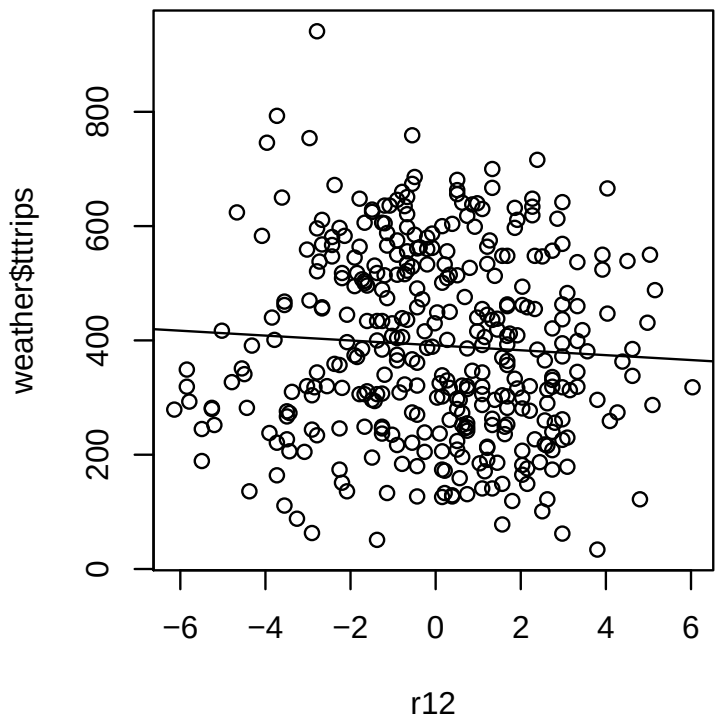




x_1 is mean temperature

$r_{1|2}$ is mean temperature "adjusting for max temperature"





```r
fit_x1x2<-lm(weather$tttrips~x1+x2) ; summary(fit_x1x2)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -251.829061  31.773619 -7.925728 2.851907e-14
## x1            -4.274695   2.381626 -1.794864 7.351178e-02
## x2            13.601363   2.013388  6.755459 5.717266e-11
```

```r
fit_x1x2<-lm(weather$tttrips~x1+x2) ; summary(fit_x1x2)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -251.829061  31.773619 -7.925728 2.851907e-14
## x1            -4.274695   2.381626 -1.794864 7.351178e-02
## x2            13.601363   2.013388  6.755459 5.717266e-11
```

```r
fit_x2r12<-lm(weather$tttrips~x2+r12) ; summary(fit_x2r12)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -268.787462  30.336326 -8.860251 3.688357e-17
## x2            10.081213   0.455292 22.142302 3.116442e-69
## r12           -4.274695   2.381626 -1.794864 7.351178e-02
```

```r
fit_x1x2<-lm(weather$tttrips~x1+x2) ; summary(fit_x1x2)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -251.829061  31.773619 -7.925728 2.851907e-14
## x1            -4.274695   2.381626 -1.794864 7.351178e-02
## x2            13.601363   2.013388  6.755459 5.717266e-11
```

```r
fit_x2r12<-lm(weather$tttrips~x2+r12) ; summary(fit_x2r12)$coef
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -268.787462  30.336326 -8.860251 3.688357e-17
## x2            10.081213   0.455292 22.142302 3.116442e-69
## r12           -4.274695   2.381626 -1.794864 7.351178e-02
```

```r
sum(fit_x1x2$res^2)
```

```
## [1] 4147101
```

```r
sum(fit_x2r12$res^2)
```

```
## [1] 4147101
```

Things to be aware of

We saw that adding a variable x_2 to the model can substantially

- change the estimate (even the sign) of β_j.
- change the standard error of $\hat{\beta}_j$ (and the p-value, etc.)

Things to be aware of

We saw that adding a variable x_2 to the model can substantially

- change the estimate (even the sign) of β_j.
- change the standard error of $\hat{\beta}_j$ (and the p-value, etc.)

The first phenomenon occurs when x_1 and x_2 are correlated in the sample.

Things to be aware of

We saw that adding a variable x_2 to the model can substantially

- change the estimate (even the sign) of β_j.
- change the standard error of $\hat{\beta}_j$ (and the p-value, etc.)

The first phenomenon occurs when x_1 and x_2 are correlated in the sample.

The second can occurr even when x_1 and x_2 are uncorrelated.