# Handout 6

### STAT3530 Applied Linear Models

This handout follows the lecture slides. This week we are covering the general linear model: regression with several predictors.

## Review: the simple linear model

1. In matrix form the simple linear model is $Y = X\beta + \epsilon$, with

$$\mathbb{E}[\epsilon] = _____ \qquad \text{and} \qquad \text{Var}[\epsilon] = _____$$

   With one predictor and $n$ observations, $X$ is $____ \times ____$ and $\beta$ is $____ \times ____$.

2. Fill in the following.

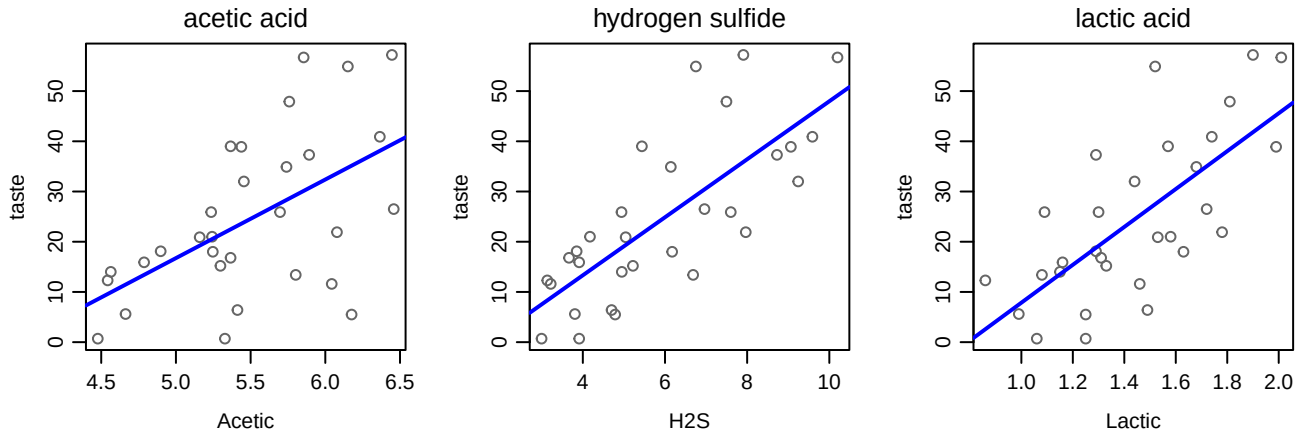
   a. The model is called *simple* because _____.

   b. In the simple fit `taste ~ H2S`, the estimated slope is 5.78. Interpret it in one sentence.

   c. Least squares gives $\widehat{\beta} =$ _____, which requires $X^TX$ to be _____.

## Several predictors at once

Thirty mature cheddars, scored by a taste panel and assayed for three compounds produced as the cheese ripens.




```
        Estimate Std. Error t value Pr(>|t|)
Acetic   15.6478     4.4958  3.4806   0.0017
H2S       5.7761     0.9458  6.1068   0.0000
Lactic   37.7199     7.1864  5.2488   0.0000
```

3. Fitted one at a time, all three compounds look significantly associated with taste. Give two reasons not to stop here.

### Predictor structure

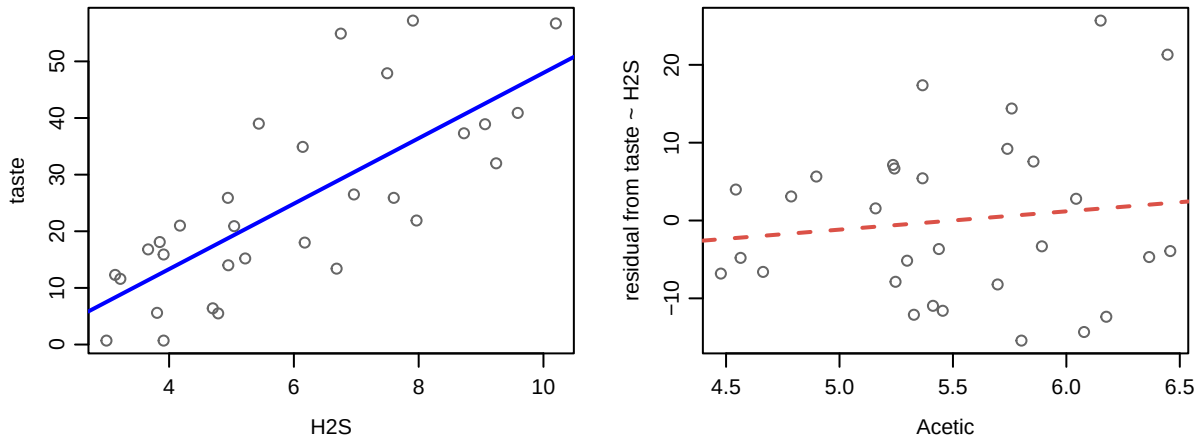
```
cor(cheddar) |> round(3)
```

```
        taste Acetic   H2S Lactic
taste   1.000  0.550 0.756  0.704
Acetic  0.550  1.000 0.618  0.604
H2S     0.756  0.618 1.000  0.645
Lactic  0.704  0.604 0.645  1.000
```

4. The predictors are correlated because _____. The most strongly correlated pair is _____, at $r =$ _____.

## Adjusting for one predictor

Below: the hydrogen sulfide model on the left, and its residuals against acetic acid on the right.




5. On its own, the acetic acid slope had $p = 0.0017$. Regressed against what is left over after fitting H2S, it has $p = 0.507$.

   a. What does the second model use as its response, and why does that change the question being asked?

   b. So the association between taste and acetic acid _____ once H2S is accounted for.

## The general linear model

6. Write the same model $Y = X\beta + \epsilon$ with $p - 1$ predictors instead of one.

   a. $X$ is _____ × _____, $\beta$ is _____ × _____, and $Y$ and $\epsilon$ are _____ × _____.

   b. The four assumptions are unchanged. Name them.

   c. Simple linear regression is the special case $p =$ _____.

7. None of the arguments from weeks 1–5 used the dimension of $X$; they only used that $X$ is _____. So fill in the results we already have:

   $\hat{\beta} =$ _____ $H =$ _____ $\hat{\beta} \sim \mathcal{N}(\beta,$ _____$)$

   and inference on a single coefficient uses

   $$T_j = \frac{\hat{\beta}_j - \beta_j}{SE(\hat{\beta}_j)} \sim t_____$$

8. What changes is the *interpretation*: $\beta_j$ is the change in mean response per unit increase in $x_j$, _____.

## Fitting and interpreting

```
fit <- lm(taste ~ Acetic + H2S + Lactic, data = cheddar)
summary(fit)$coefficients
```

```
              Estimate Std. Error    t value    Pr(>|t|)
(Intercept) -28.8767696  19.735418 -1.4631952 0.155399149
Acetic        0.3277413   4.459757  0.0734886 0.941979774
H2S           3.9118411   1.248430  3.1334077 0.004247081
Lactic       19.6705434   8.629055  2.2795710 0.031079481
```

```
confint(fit)
```

```
                2.5 %    97.5 %
(Intercept) -69.443503 11.689964
Acetic       -8.839420  9.494902
H2S           1.345656  6.478026
Lactic        1.933267 37.407820
```

9. Using the output above:

   a. The fitted model is

$$\mathbb{E}[\text{taste}] = _____ + _____ \text{Acetic} + _____ \text{H2S} + _____ \text{Lactic}$$

   b. Interpret the lactic acid coefficient in context. Be sure to say what is being held fixed.

   c. Acetic acid was significant on its own but is not here. Explain why, in one sentence.

   d. Which compounds are significantly associated with taste at the 5% level? Report a 95% interval for each and interpret one of them in context.

### Standardizing for interpretation

The predictors are on different scales, so the coefficients above cannot be compared to each other. Centering and scaling each predictor,

$$x_j \leftarrow \frac{x_j - \bar{x}_j}{s_j}$$

puts them all in standard deviations.

```
refit <- lm(taste ~ scale(Acetic) + scale(H2S) + scale(Lactic), data = cheddar)
summary(refit)$coefficients
```

```
                Estimate Std. Error    t value     Pr(>|t|)
(Intercept)   24.5333333   1.849605 13.2640910 4.413822e-13
scale(Acetic)  0.1871004   2.545979  0.0734886 9.419798e-01
scale(H2S)     8.3200133   2.655260  3.1334077 4.247081e-03
scale(Lactic)  5.9698141   2.618832  2.2795710 3.107948e-02
```

```
confint(refit)
```

```
                   2.5 %    97.5 %
(Intercept)   20.7314151 28.335252
scale(Acetic) -5.0462334  5.420434
scale(H2S)     2.8620474 13.777979
scale(Lactic)  0.5867273 11.352901
```

10. Using the output above:

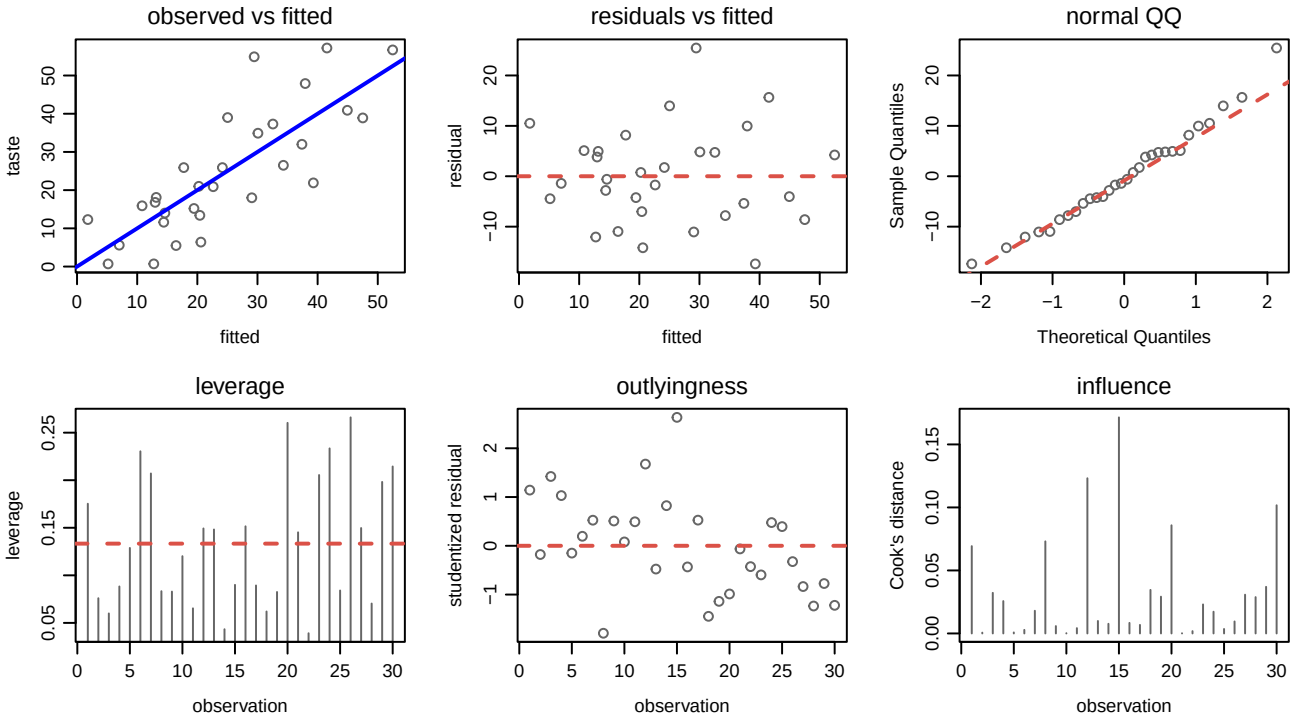
    a. After standardizing, a one unit increase in a predictor means _____.

    b. A +1 SD change in H2S is associated with a change in mean taste of between _____ and _____ points; for lactic acid, between _____ and _____ points. Which compound matters more, and by how much?

    c. Compare the $t$ values and $p$-values to the unstandardized fit. What happened to them, and why?

## Diagnostics and specification

The same six plots as before, for the cheddar model.

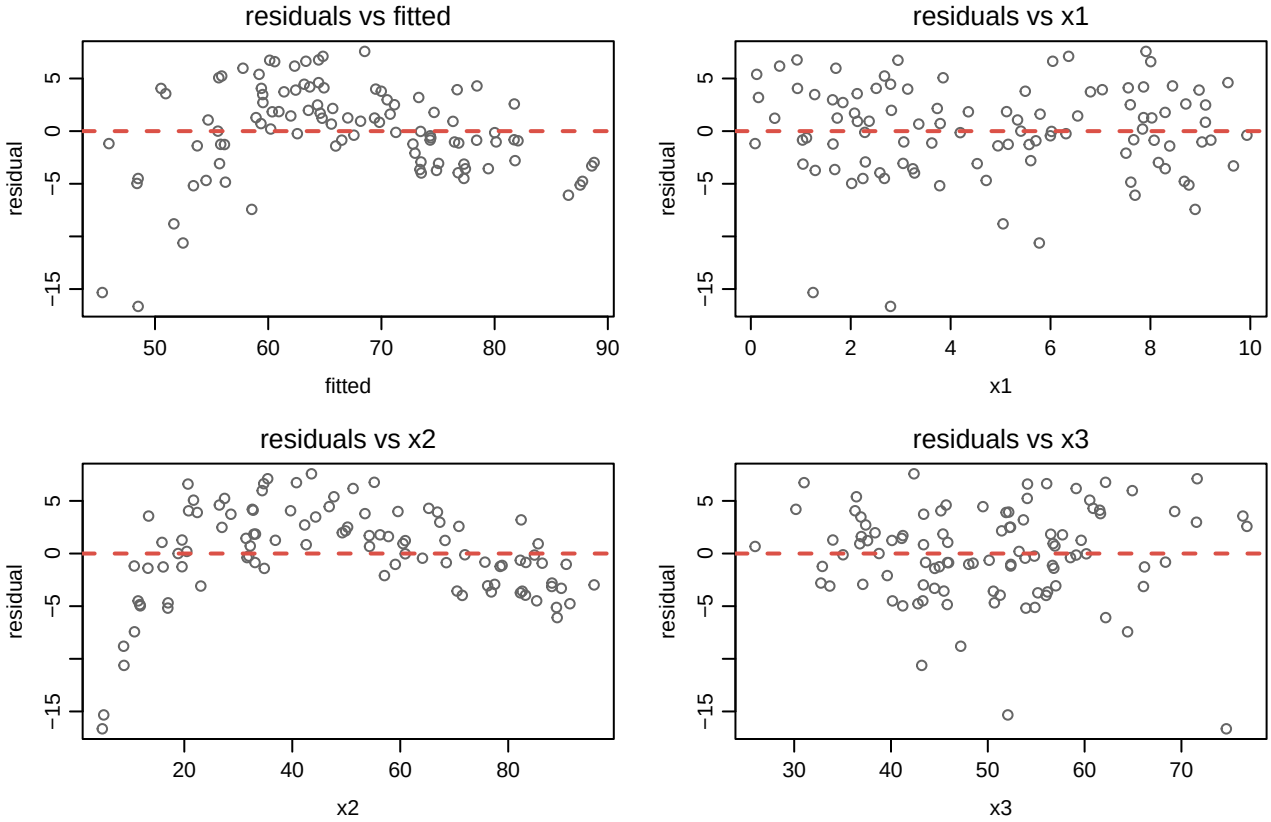



11. Fill in what each plot checks, and what it says about this model.

| plot | checks | verdict here |
|---|---|---|
| residuals vs fitted | | |
| normal QQ | | |
| leverage | | |
| studentized residuals | | |
| Cook's distance | | |

### What if you *had* spotted an issue?

Below, data simulated from $Y = \beta_0 + \beta_1 x_1 + \beta_2 \log(x_2) + \beta_3 x_3 + \epsilon$, fit with all three predictors untransformed.




12. Using the plots above:

    a. The residuals vs fitted plot shows _____.

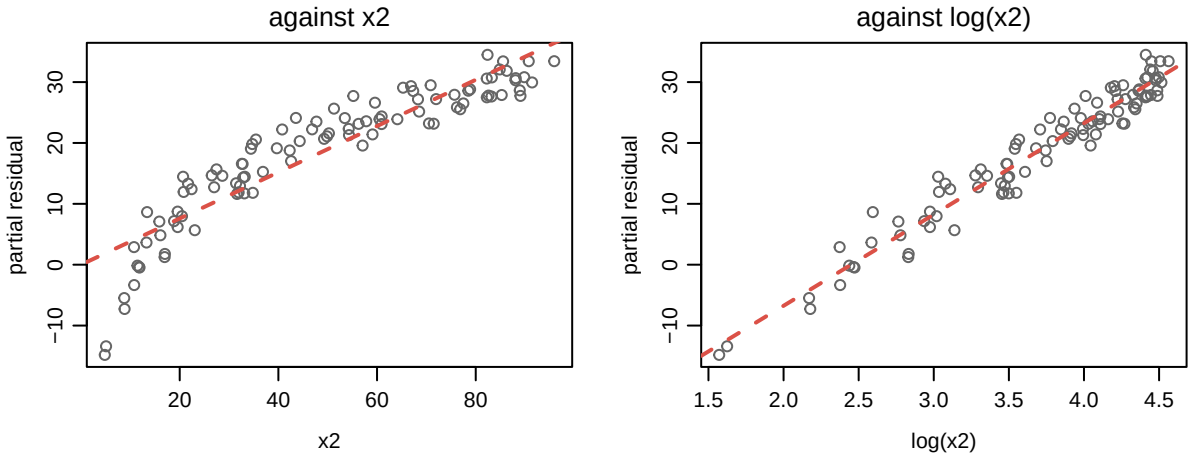
    b. Why can't that plot tell you *which* predictor is at fault?

    c. Which plot does identify the culprit, and which predictor is it?

## Partial residuals

A residual plot shows only what is left over. Adding back one predictor's fitted contribution shows the shape that predictor's effect should have:

$$e_j = _____ + _____$$




13. Using the plots above:

    a. What does a partial residual plot show that a plain residual plot does not?

    b. The left plot traces a _____, so the fix is _____.

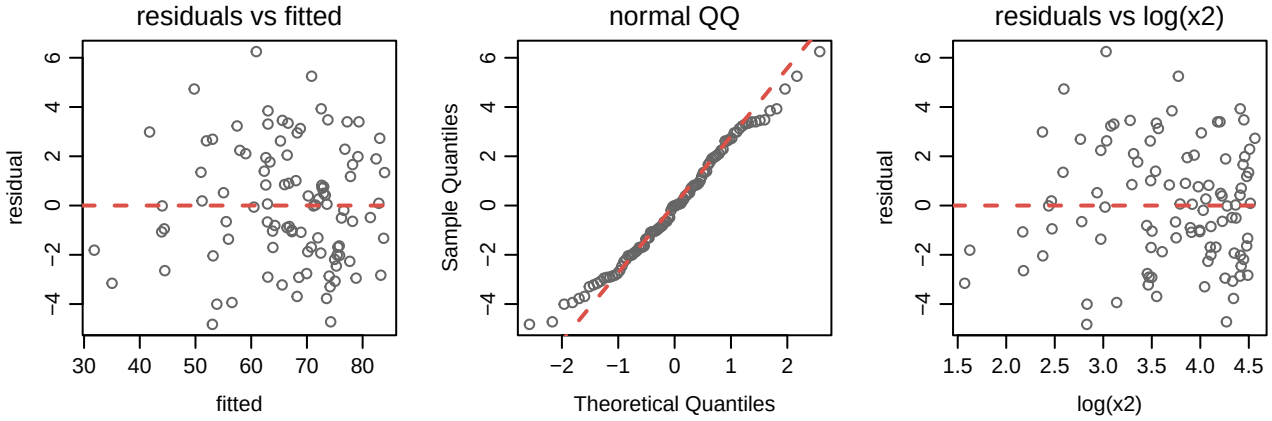
### The corrected model

```
better <- lm(y ~ x1 + log(x2) + x3, data = sim3)
summary(better)$coefficients
```

```
               Estimate Std. Error    t value     Pr(>|t|)
(Intercept)  0.92417884 1.82191177  0.5072577 6.131373e-01
x1           1.15383750 0.08595770 13.4233168 9.364092e-24
log(x2)     15.01377350 0.34633218 43.3507892 7.568222e-65
x3           0.09746968 0.02244789  4.3420414 3.499016e-05
```

```
c(bad = summary(bad)$r.squared, better = summary(better)$r.squared) |> round(3)
```

```
   bad better
 0.849  0.955
```




14. Using the output and plots above:

    a. Are the diagnostics clean now? $R^2$ rises from _____ to _____.

    b. Interpret the $\log(x_2)$ coefficient. What does a doubling of $x_2$ do to the mean response?

## Prediction

```
new <- data.frame(Acetic = c(5.0, 6.2), H2S = c(4.0, 9.0), Lactic = c(1.2, 1.8))

predict(full, new, interval = 'confidence')
```

```
       fit       lwr      upr
1 12.01395  6.578536 17.44937
2 43.76877 36.786956 50.75059
```

```
predict(full, new, interval = 'prediction')
```

```
       fit       lwr     upr
1 12.01395 -9.507692 33.5356
2 43.76877 21.805549 65.7320
```

15. Using the output above:

    a. Write the row of $x_0$ used for the first new cheese.

    $$x_0 = [\ _____ \quad _____ \quad _____ \quad _____\ ]$$

    b. Fill in the two standard errors and say which interval is wider, and why.

    $$SE(\hat{Y}) = \hat{\sigma}\sqrt{\qquad\qquad} \qquad\qquad SE(\hat{Y}_{\text{pred}}) = \hat{\sigma}\sqrt{\qquad\qquad}$$

    c. For the second cheese, report the predicted taste score and the interval you would give a cheesemaker who wants to know how *this one cheese* will score.

## When the columns are not independent

Recall that $X$ must be full rank, otherwise $X^TX$ is singular and $\hat{\beta}$ cannot be computed.

16. Full rank for $X_{n\times p}$ requires that the columns are _____ and that $n$ _____ $p$.

```
cheddar3 <- cheddar
cheddar3$Lactic2 <- 2 * cheddar$Lactic
lm(taste ~ ., data = cheddar3)
```

```
Call:
lm(formula = taste ~ ., data = cheddar3)

Coefficients:
(Intercept)       Acetic           H2S       Lactic      Lactic2
   -28.8768        0.3277       3.9118      19.6705           NA
```

17. What did R do with the fourth predictor, and why was it the only option?

### Multicollinearity

Predictors are rarely *exactly* dependent, but they can come close. Here is a new compound that tracks lactic acid almost perfectly.

```
set.seed(5)
cheddar2$Schmactic <- cheddar$Lactic + rnorm(30, mean = 0, sd = 0.05)
fit.multi <- lm(taste ~ ., data = cheddar2)
summary(fit.multi)$coefficients
```

```
               Estimate Std. Error    t value    Pr(>|t|)
(Intercept) -26.1252632  20.060985 -1.3022922 0.204687213
Acetic       -0.4758645   4.569705 -0.1041346 0.917893442
H2S           4.0140556   1.258994  3.1883044 0.003823958
Lactic       54.8702045  40.720624  1.3474795 0.189912047
Schmactic   -34.4514830  38.942246 -0.8846815 0.384758262
```

18. Compare this output to the fit in question 9.

    a. The standard error on lactic acid went from _____ to _____, and lactic acid is now _____.

    b. Is this a failure of the model, or a correct report? Explain.

19. Suppose two predictors are related by $x_1 \approx a\,x_2 + b$. Complete the algebra:

$$\mathbb{E}[Y] \approx \beta_0 + \beta_1 x_1 + \beta_2(a\,x_1 + b) = _____ + _____\; x_1$$

Since $a$ and $b$ are unknown, different combinations of $\beta_1$ and $\beta_2$ give _____ fits.

20. Geometrically:

    a. With one predictor, exact collinearity means $x = c\mathbf{1}$ and $col(X)$ is _____ instead of a plane.

    b. With two correlated predictors, $col(X)$ is still a plane, but _____.

    c. In both cases the projection $\hat{Y}$ is _____; what suffers is _____.

21. Mathematically, near collinearity drives $\det(X^TX)$ toward _____, so the entries of $(X^TX)^{-1}$ get _____, and since $\mathrm{Var}[\hat{\beta}] =$ _____, the standard errors get _____.

### A real example

`hipcenter` is how far a driver's hips sit from the steering wheel, for 38 drivers, along with eight body measurements.

```
seatpos <- faraway::seatpos
seat <- lm(hipcenter ~ ., data = seatpos)
round(summary(seat)$coefficients, 3)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  436.432    166.572   2.620    0.014
Age            0.776      0.570   1.360    0.184
Weight         0.026      0.331   0.080    0.937
HtShoes       -2.692      9.753  -0.276    0.784
Ht             0.601     10.130   0.059    0.953
Seated         0.534      3.762   0.142    0.888
Arm           -1.328      3.900  -0.341    0.736
Thigh         -1.143      2.660  -0.430    0.671
Leg           -6.439      4.714  -1.366    0.182
```

```
cor(seatpos$Ht, seatpos$HtShoes)
```

```
[1] 0.9981475
```

```
round(summary(lm(hipcenter ~ Ht, data = seatpos))$coefficients, 3)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  556.255     90.670   6.135        0
Ht            -4.265      0.535  -7.970        0
```

22. The full model has $R^2 = 0.687$.

    a. How many of the eight predictors are significant at the 5% level? _____

    b. The `Ht` coefficient is $+0.60$ in the full model and $-4.27$ on its own. Which would you report, and why?

    c. A designer asks how much further back a taller driver sits. Can this data answer that? Can it answer *which* body measurement matters?

## Activity: your turn

In 1971 a sample of 102 Canadian occupations was rated for prestige. The data pair those ratings with the average **education** (years), average **income** (dollars), and percentage **women** in each occupation.

**Which of the three are associated with prestige, and how much is each worth?**

The script `scripts/09-multiple-regression.R` walks through the same five steps on the cheddar data first.

1. Look at the data with `pairs()`. What looks straight? What does not?

2. Fit a model and read `summary()`.

3. Check the diagnostics, including one residual plot per predictor.

4. Check the partial residual plots. Refit if a predictor is on the wrong scale.

5. Report confidence intervals and answer the question above. If you transformed a predictor, say what its coefficient means on the original scale.