Activity #12: More regression topics: LOWESS; polynomial, nonlinear, robust, quantile; ANOVA as regression

Scenario: 31 counts (over a 30-second period) were recorded from a Geiger counter at a nuclear plant:

```
                Time      Count (above background levels)
                0         126.6
                1         101.8
                2          71.6
                …           …
                30         19.3
                =============
        Mean =  15        43.745
        SD =    9.09      29.308
        N = 31
        r = -0.877
```

1. Let's model the counts as a function of time. To begin, we can fit a simple linear model. Based on the R-squared value of 0.7687, we might feel satisfied that this linear model adequately describes the data. Based on the residual plots, evaluate whether the necessary assumptions were satisfied for this linear regression analysis:

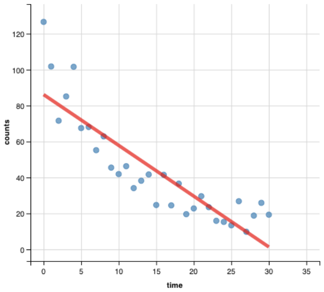
Model: count = b_0 + b_1(time)

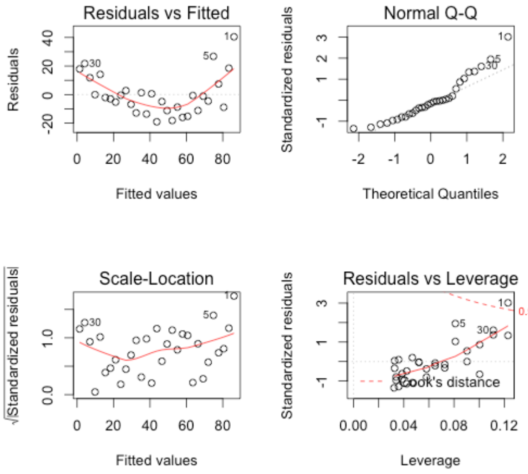
Least-squares line: y = 86.14 – 2.826x

R^2 = 0.7687

RMSE = 14.34

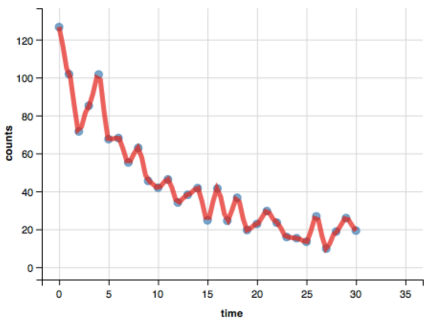
F = 96.397 (p = 9.991e-11)



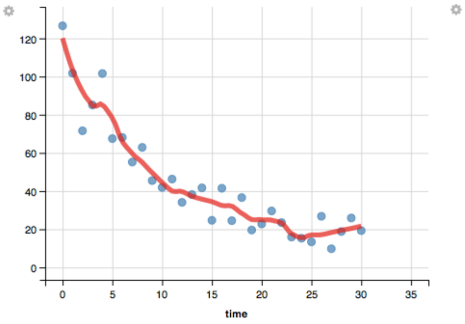


2. It looks like we have a linearity issue (which would have been obvious had we plotted the data prior to fitting the linear model). To get a better idea of the "shape" of our scatterplot, we can use *locally weighted scatterplot smoothing* (LOWESS). LOWESS connects a series of models fit to small (local) subsets of the data (defined by the bandwidth or span) to describe the general relationship.

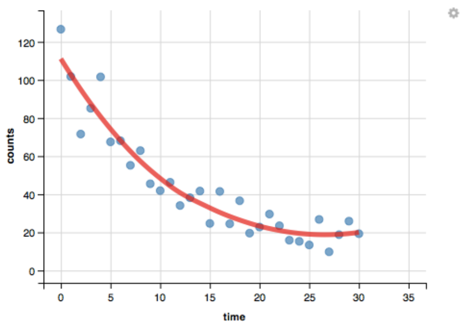
To see an animation of LOWESS in action, go to: http://bradthiessen.com/html5/stats/m301/lowess.gif



Bandwidth = 0.1



Bandwidth = 0.3



Bandwidth = 1.0

3. Based on that LOWESS curve, it's obvious that a linear function isn't the best fit for our data. So far in this course, we've only fit linear models to data. We've visualized this as a straight line (or plane) through a scatterplot.

   What can we do when our data appear to have a "curved" relationship? Believe it or not, we can still use linear regression. Linear regression can include "lines" that have curved shapes.

   One way to do this is to use polynomial (or curvilinear) regression to fit a curve through the data. In this Geiger counter example, we could try to fit the following models:

$$\text{Linear model: } \hat{y} = b_0 + b_1(\text{time})$$

$$\text{Model with quadratic term: } \hat{y} = b_0 + b_1(\text{time}) + b_2(\text{time})^2$$

$$\text{Model with cubic term: } \hat{y} = b_0 + b_1(\text{time}) + b_2(\text{time})^2 + b_3(\text{time})^3$$

   These are all considered to be linear models? Why? If you take the partial derivative of your model and the result no longer includes the unknown coefficients, then the model is considered to be linear. If, on the other hand, the partial derivative results in a function that does still include the unknown coefficients, then the model is considered to be nonlinear. Let's take a look at the partial derivative of the model with the cubic term:

$$\frac{dy}{db_0} = 1 \qquad \frac{dy}{db_1} = x \qquad \frac{dy}{db_2} = x^2 \qquad \frac{dy}{db_3} = x^3$$

   Since these partial derivatives are no longer functions of the coefficients (b values), this is a linear model. It is a linear combination of predictor variables.

   We can fit the models with the quadratic and cubic terms and compare them to the null model with no predictors:

Model: count = b_0 + b_1(time)

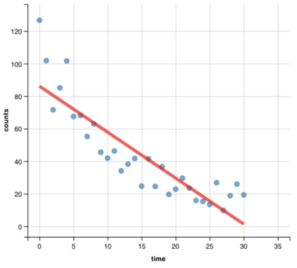
Formula: $y = 86.1 – 2.83x$

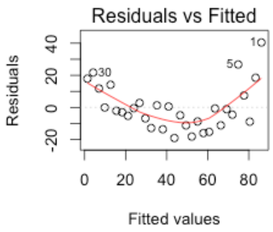
$R^2 = 0.7687$

RMSE = 14.34

Compared to null model:

F = 96.397 (p = 9.991e-11)





Model: y = b_0 + b_1(time) + b_2(time)^2

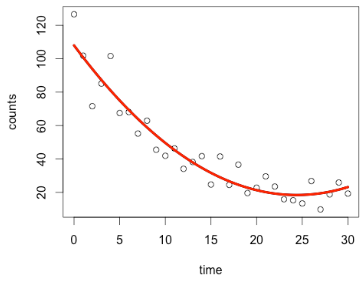
Formula: $y = 108 – 7.34x + 0.15x^2$

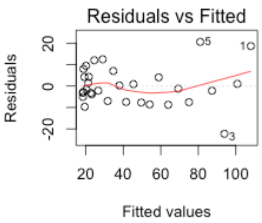
$R^2 = 0.9079$

RMSE = 9.205

Compared to null model:

F = 138.1 (p = 3.143e-15)





Model: y = b_0 + b_1(time) + b_2(time)^2 + b_3(time)^3

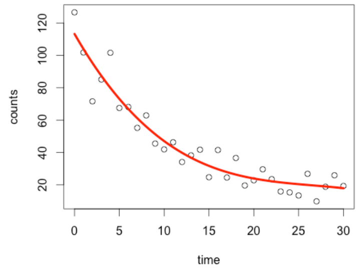
Formula: $y = 113.3 – 9.65x + 0.35x^2 – 0.004x^3$

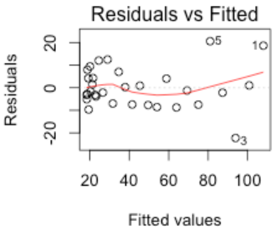
$R^2 = 0.915$

RMSE = 9.007

Compared to null model:

F = 96.88 (p = 1.442e-14)





4. We can compare these three models using the omnibus F-test. Verify the calculations and interpret. From this, which model would you conclude fits the data the best?

```
Analysis of Variance Table

Model 1: counts ~ 1
Model 2: counts ~ time
Model 3: counts ~ time + I(time^2)
Model 4: counts ~ time + I(time^2) + I(time^3)

  Res.Df     RSS Df Sum of Sq       F    Pr(>F)
1     30 25769.0
2     29  5959.5  1   19809.5 244.176 4.763e-15
3     28  2372.4  1    3587.1  44.215 3.897e-07
4     27  2190.5  1     182.0   2.243    0.1458
```

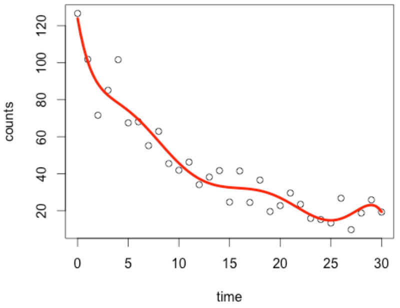
5. Look at the R-squared values for each model. What do you think would happen to the R-squared value if we fit a model with higher-powered terms (such as including all the way up to the 7th power)? Why?

6. Let's go ahead and fit the model that includes the 7th power:

Formula: $y = 124 - 31.6x + 9.9x^2 - 1.7x^3 + 0.15x^4 - 0.007x^5 + 0.00017x^6 - 0.0000016x^7$

$R^2 = 0.9314$

RMSE = 9.007



If we compare it to the model with the quadratic term, we find:

```
Analysis of Variance Table

Model 1: counts ~ time + I(time^2)
Model 2: counts ~ time + I(time^2) + I(time^3) + … + I(time^6) + I(time^7)

  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     28 2372.4
2     23 1767.2  5    605.22 1.5754 0.2066
```

From this, what can we conclude?

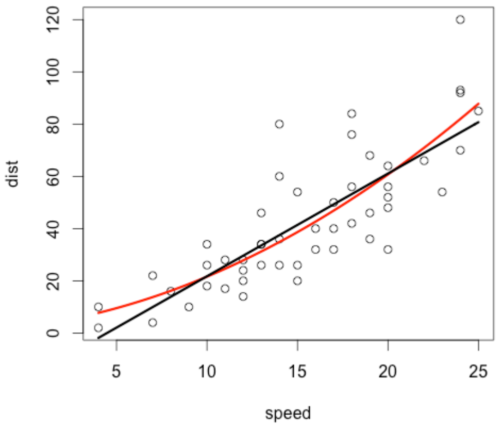
7. Just like in the last activity, we can use cross-validation to help choose the best model:

Average cross-validated mean square error:	Model
219	count = f(time)
103	count = f(time + time^2)
105	count = f(time + time^2 + time^3)
110	count = f(time + time^2 + time^3 + time^4)

From this, what can we conclude?

8. Let's look at another quick example. I had a dataset containing the speed and stopping distance of 50 cars. I fit models including linear, quadratic, and cubic terms and then tested each model in order:

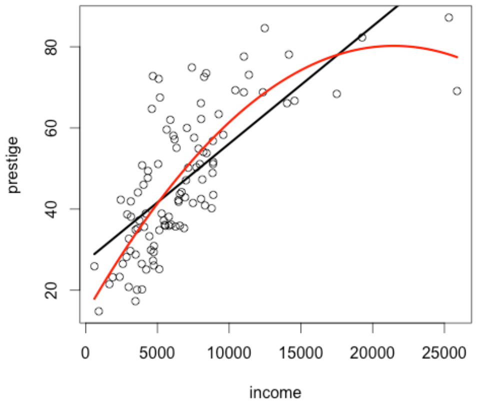
```
Analysis of Variance Table

Model 1: dist ~ 1
Model 2: dist ~ speed
Model 3: dist ~ speed + I(speed^2)
Model 4: dist ~ speed + I(speed^2) + I(speed^3)

  Res.Df   RSS Df Sum of Sq       F    Pr(>F)
1     49 32539
2     48 11354  1   21185.5 91.6398 1.601e-12
3     47 10825  1     528.8  2.2874    0.1373
4     46 10634  1     190.4  0.8234    0.3689
```



Which model would you choose to describe this data?

9. Finally, let's go back to our "prestigious occupations" dataset. We can fit various models of prestige as a function of income. Based on this output (and the cross-validation output on the next page), which model would you choose?

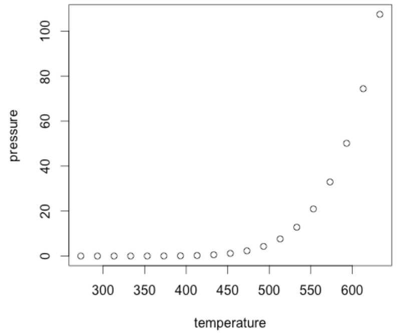
```
Analysis of Variance Table

Model 1: prestige ~ 1
Model 2: prestige ~ income
Model 3: prestige ~ income + I(income^2)
Model 4: prestige ~ income + I(income^2) + I(income^3)

  Res.Df   RSS Df Sum of Sq        F    Pr(>F)
1    101 29895
2    100 14616  1   15279.3 124.0617 < 2.2e-16 ***
3     99 12077  1    2539.3  20.6179 1.597e-05 ***
4     98 12070  1       7.4   0.0598    0.8073
```



Average cross-validated mean square error:	Model
154	prestige = f(income)
130	prestige = f(income + income^2)
131	prestige = f(income + income^2 + income^3)
134	prestige = f(income + income^2 + income^3 + income^4)

Scenario: We're going to investigate a dataset containing 19 observations:

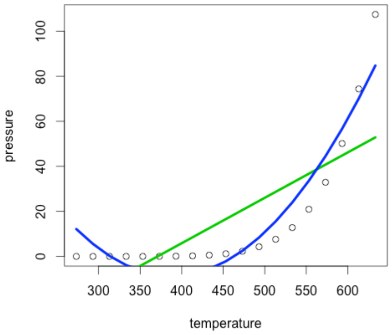
- temperature in Kelvin
- vapor pressure of mercury in kiloPascals



Source: Handbook of Chemistry and Physics, CRC Press (1973)

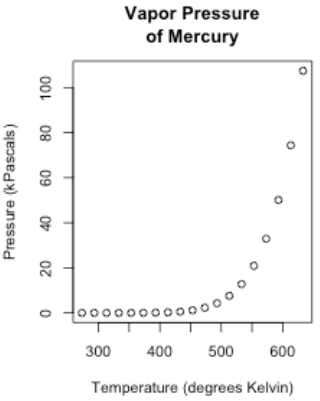
10. We can try to fit linear and quadratic models to this data. The plot shows neither model fits well, though.

Model: pressure = $b_0 + b_1$(temp)
R^2 = 0.5742
RMSE = 20.1
Compared to null model:
F = 22.93 (p = 0.000171)

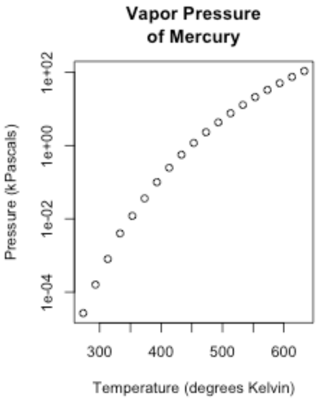


Model: pressure = $b_0 + b_1$(temp) + b_2(temp)^2
R^2 = 0.9024
RMSE = 9.92
Compared to null model:
F = 74 (p = 8.209e-09)

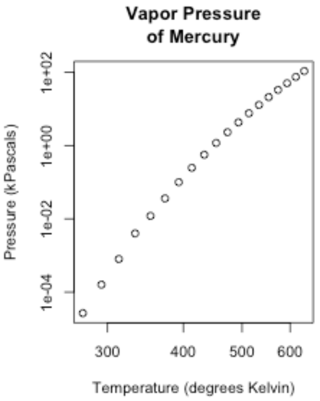
We could also try to fit a nonlinear model to this data. The scatterplot indicates some kind of exponential model would fit the data. We could transform our variables to get an idea about which model might fit best.



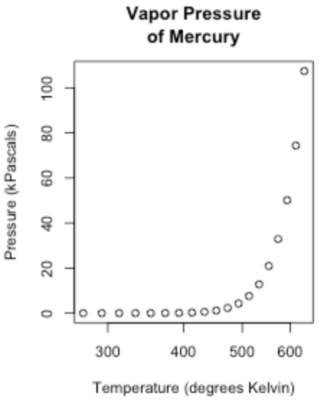
(no transformation)



log(pressure)



log(pressure)
and log(temp)



log(temp)

11. Based on the graphs, it looks like a natural log transformation of both the pressure and temperature variables resulted in data that are approximately linear. This means we might want to try to fit one of these (identical) power models:

$$\hat{y} = b_0\left(\text{temp}\right)^{b_1}$$

$$\ln\left(\hat{y}\right) = b_0 + b_1 \ln\left(\text{temp}\right)$$

If we use the second parameterization, we can still use our ordinary least squares methods to find the line of best fit. I used R to obtain the following output:

Model: log(pressure) = b_0 + b_1*log(temp)

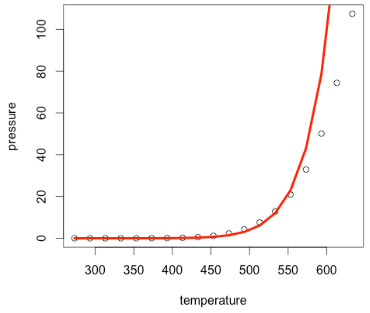
Formula: log(y) = -108.114 + 17.615log(x)

R^2 = 0.9853

RMSE = $e^{0.5736}$ = 1.77

Compared to null model:

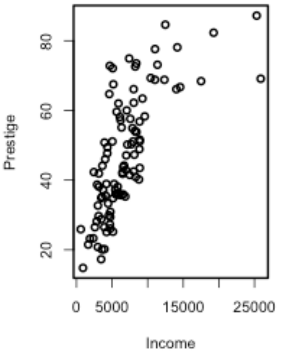
F = 1142 (p < = 2.2e-16)



Because this power model predicts log(pressure), we cannot use the omnibus F-test to compare it to our linear and quadratic models.

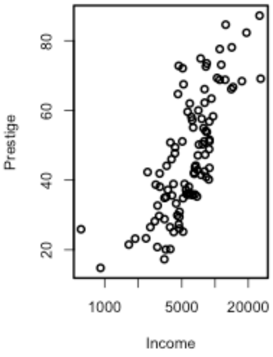
Based on this output, which model would you choose to predict pressure based on temperature?

12. We can replicate this nonlinear regression process with our prestige data. Below, I've graphed prestige as a function of income with some transformations to the variables:



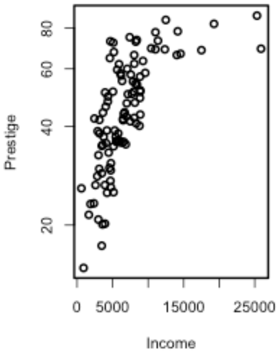
(no transformation)

linear model



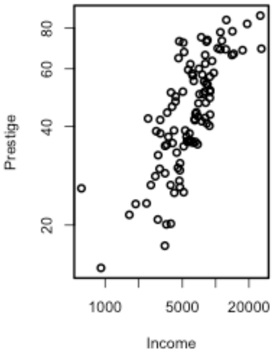
log(prestige)

log(y) model



log(prestige)
and log(income)

power model



log(income)

log(x) model

Based on these plots, which transformation yielded the most "linear-looking" scatterplot?

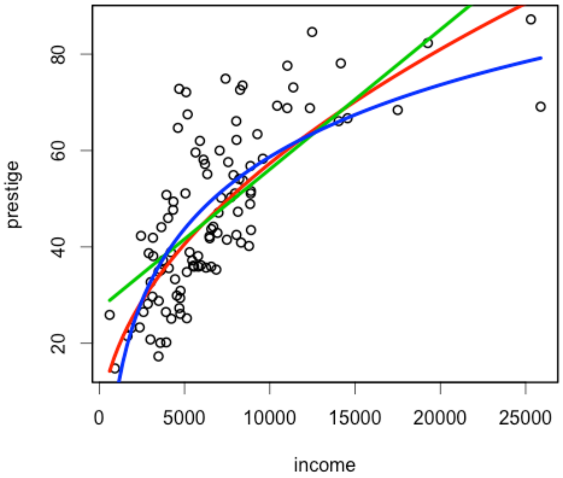
13. I then fit these models to the data an obtained the following:

Linear Model: prestige) = b_0 + b_1(income)
R^2 = 0.511

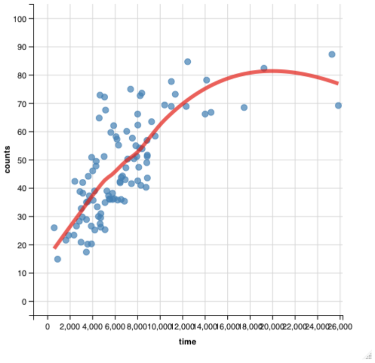
Log(y) model: log(prestige) = b_0 + b_1(income)
R^2 = 0.452

Power model: log(prestige) = b_0 + b_1*log(income)
R^2 = 0.566

log(x) model: prestige = b_0 + b_1*log(income)
R^2 = 0.549



We could then compare all these models to the LOWESS curve:



For another nonlinear regression example (including other function options), go to:
http://bradthiessen.com/html5/stats/m301/16e.pdf

14. Let's take a step back and take a look at all the regression-related techniques we've covered thus far:
    - Simple linear regression
    - Diagnostic plots
    - Tests for significance of coefficients (randomization-based, t-test for correlation, t-test for slope, F-test)
    - Confidence intervals and prediction intervals
    - Multiple linear regression
    - Model comparison (using omnibus F-test)
    - Adjusted R-squared
    - Standardized beta coefficients
    - Interaction terms
    - Model selection (forward, backward, stepwise)
    - Cross-validation
    - Best subsets regression
    - Ridge regression / Lasso
    - LOWESS
    - Polynomial (curvilinear) regression
    - Nonlinear regression
    - Robust regression
    - Quantile regression
    - ANOVA as regression

} On the next page, we'll learn some aspects of these techniques

15. As we've discussed, the linear and nonlinear regression techniques we've covered thus far are all based on some underlying assumptions. We've investigated some techniques (visualizations and tests) to evaluate these assumptions. What can we do if we find the assumptions are violated?

    If we have an issue with normality of residuals, we might choose to transform our variables. Likewise, if our linearity assumption is violated, we might turn to a curvilinear or nonlinear regression model. But what do we do if we have concerns about the homogeneity of variances assumption? Likewise, what can we do if our data have a few influential outliers?

    One thing we could do is run a *robust* regression. There are two types of robust regression methods:

    - Regression with robust standard errors
    - Robust estimation of coefficients and standard errors

    For a couple quick examples of robust regression methods, check out:
    http://www.philender.com/courses/linearmodels/notes4/robust.html

    We'll briefly investigate one approach to robust regression using *bootstrap* methods. In this approach, we:

    a) Estimate regression coefficients from the data: $\hat{y} = b_0 + b_1 x_1 + \ldots$

    b) Calculate predicted values $(\hat{y})$ and residuals for each observation $(e = \hat{y} - y)$

    c) Take all *n* residuals and select a sample of *n* of them with replacement (the bootstrap sample).

    d) Using those sampled residuals, calculate new Y values $(y^* = \hat{y} + e)$

    e) Now run a regression using your original x variable and the new y* variable values

    f) Repeat steps c-e many times (say, 10,000 times)

    Now that we have 10,000 estimates of our regression coefficients, we can estimate their standard errors by simply calculating the standard deviation of all our coefficient estimates. Likewise, we can find the lowest and highest 2.5% of the coefficient estimates to estimate a 95% confidence interval for each.

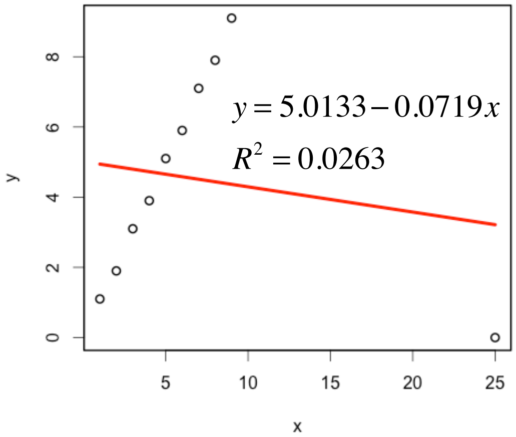
    To read about this technique, check out either of the following:
    http://socserv.mcmaster.ca/jfox/Books/Companion/appendix/Appendix-Bootstrapping.pdf
    http://www.sagepub.com/upm-data/21122_Chapter_21.pdf

    Let's try this out on our prestige data. We'll model prestige as a function of income, education, and %women:

	Linear Model		**Bootstrap Method**	
	Coefficient	Standard Error	Coefficient	Standard Error
(intercept)	-6.794334	3.239089	-6.7918	3.24658
Income	0.001314	0.000278	0.00136	0.00028
Education	4.186637	0.388701	4.1829	0.39107
%women	-0.008905	0.030407	-0.00981	0.03072

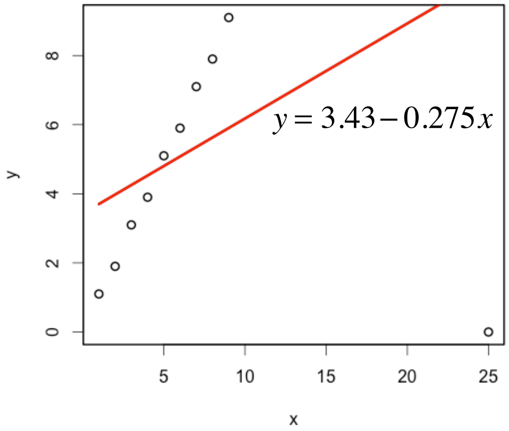
    Notice the bootstrap coefficients are slightly biased (compared to our original model). Also, notice the bootstrap standard errors are all larger. What's the consequence of having larger standard errors for the coefficients?

16. Let's take a different approach to bootstrapping a regression. We'll start with a fictitious dataset that has a clear outlier. When we fit a linear model to the data, we get what's pictured to the right.



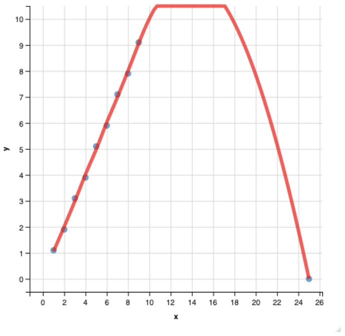
That outlier on the bottom-right obviously had a significant impact on our estimates of the regression coefficients. To mitigate the effect of that outlier, we could choose to:

a) Take a random sample of n observations <u>with replacement</u> from our data.
b) Estimate the regression coefficients for this bootstrap sample.
c) Repeat this process many times to end up with lots of estimated coefficients
d) Use the mean (or median) of those bootstrap coefficients



The bootstrap regression line is pictured to the right. While it's still not a great fit, it's markedly better than the original.

Just for comparison, a lowess curve is displayed below:



Let's see how this bootstrap method works on our prestige dataset. Below, I've pasted the coefficients using ordinary least squares and using the bootstrap method. Once again, notice that the coefficients and standard errors differ slightly.

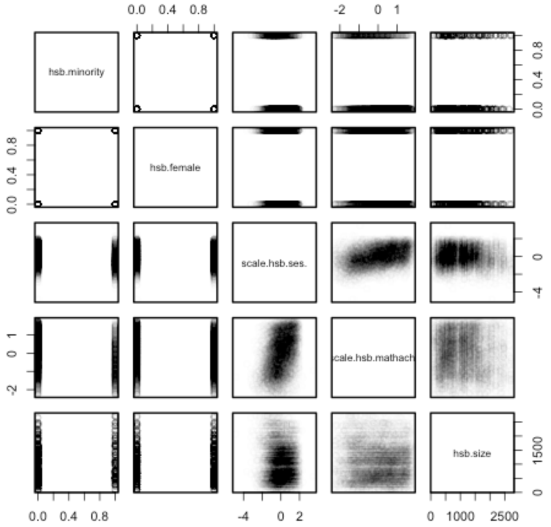
	Linear Model		**Bootstrap Method**	
	<u>Coefficient</u>	<u>Standard Error</u>	<u>Coefficient</u>	<u>Standard Error</u>
(intercept)	-6.794334	3.239089	-6.71	3.1224
Income	0.001314	0.000278	0.00142	0.0005
Education	4.186637	0.388701	4.10	0.4673
%women	-0.008905	0.030407	-0.00466	0.0370

What assumptions did we make with the ordinary least squares regression? What assumptions did we make using the bootstrap method?

Scenario: The National Center for Education Statistics (NCES) is mandated to "collect and disseminate statistics and other data related to education in the United States." To this end, it has initiated several large scale studies in which a cohort is studied at regular intervals over several years. The High School and Beyond (HSB) study tracked achievement and social aspects of the 1980 sophomore & senior classes.

The dataset we have consists of 7,185 students from 160 different high schools. The following variables were measured:

- schid: school ID number
- minority: 0 = no; 1 = yes
- female: 0 = no; 1 = yes
- ses: socioeconomic status of the student (z-score)
- **mathach: math achievement score (z-score)**
- size: number of students in school
- schtype: school type (0 = public; 1 = private)
- meases: socioeconomic status of the school



We'll focus on **mathach** as our outcome variable.

Data: http://www.bradthiessen.com/html5/data/hsb.csv

17. From the scatterplots, we can see math achievement is associated with socioeconomic status. We can fit the following linear model:

Model: mathach = b_0 + b_1(ses)

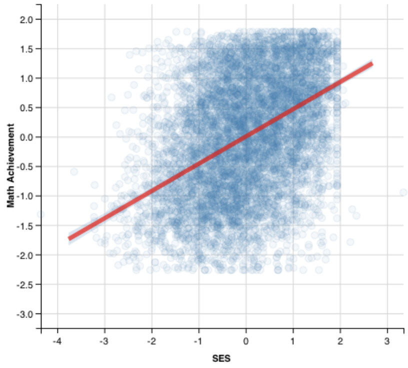
Formula: y =0 + 0.361x

R^2 = 0.13

RMSE = 0.933

Comparison to null model:

F = 1075 (p < 2e-16)



Take a look at the p-value and R-squared value. What can we say about the relationship between SES and math achievement? Keep in mind our variables are standardized (z-scores).

18. When we fit an ordinary least squares regression line, we're fitting a single line through the scatterplot. We can add a confidence or prediction interval, but we're assuming the relationship between our variables is virtually the same for all observations in the data. So, in this example, we're assuming a 1 standard deviation increase in SES is associated with a 0.361 standard deviation increase in math achievement for all students (regardless of just how high or low the students math achievement is).

Might we expect SES to have more or less of an impact on extremely low- or high-achieving students? Maybe. To investigate this, we can use **quantile regression**. What is a quantile?

19. The median is the 2-quantile (50th percentile). We could run a quantile regression for the 50th percentile.

Linear Regression

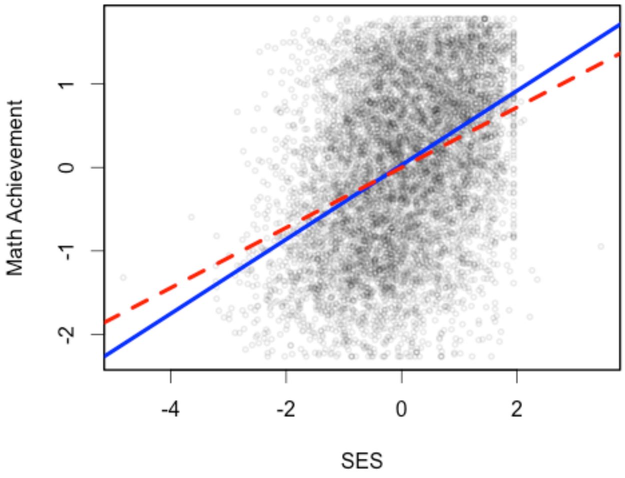
Model: mathach = $b_0 + b_1$(ses)

Formula: $y = 0 + 0.361x$

Quantile Regression for median

Model: median(mathach) = $b_0 + b_1$(ses)

Formula: $y = 0.0295 + 0.4454x$

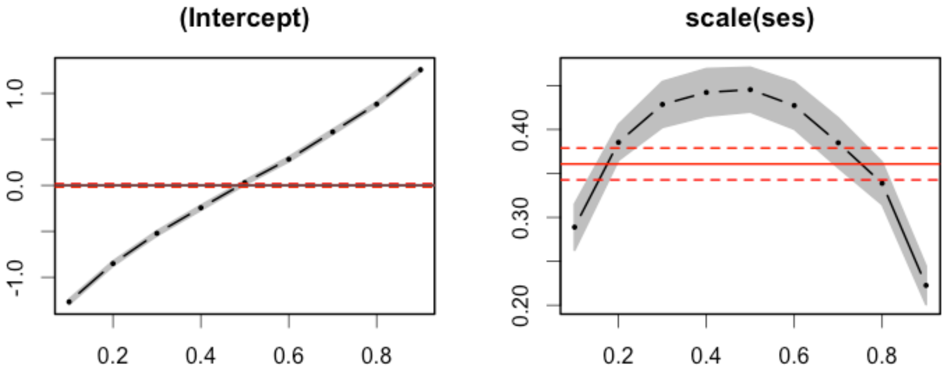


Interpret the coefficients of this quantile regression line. Remember that these coefficients only apply to the chosen quantile of interest!

20. That median regression can be useful (especially when dealing with outliers), but we're more interested in determining if the relationship between SES and math achievement differ depending on the level of math achievement.

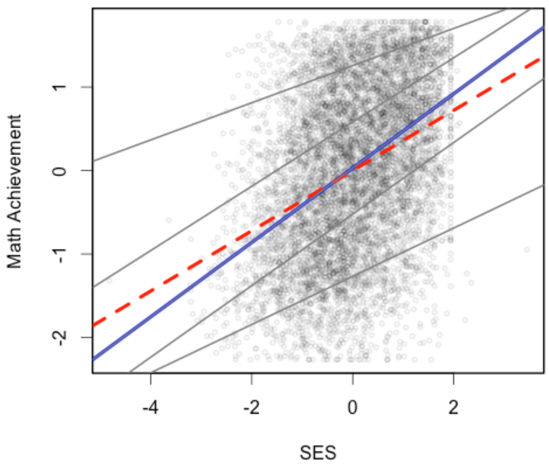
To investigate this, we can run a quantile regression for as many percentiles as we want. For example, we could run the quantile regression for the 10th, 20th, 30th, …, 80th, and 90th percentiles. That would show us whether the relation between SES and math achievement changes across almost the entire distribution of math achievement scores.

Rather than listing out the coefficients for all 9 regression model, we can display the magnitude of the coefficients across each of the 10 deciles:



The red line shows the coefficients for our simple linear regression model (along with confidence bands). The black lines show the coefficient estimate across different percentiles of math achievement (along with grey confidence bands). From this, what can we conclude about the relationship between SES and math achievement?

21. We can also display these quantile regression lines on the scatterplot. To the right, I've plotted the regression lines for the 10th, 30th, 50th, 70th, and 90th percentiles of math achievement. Do these results match the graphs on the bottom of the previous page?



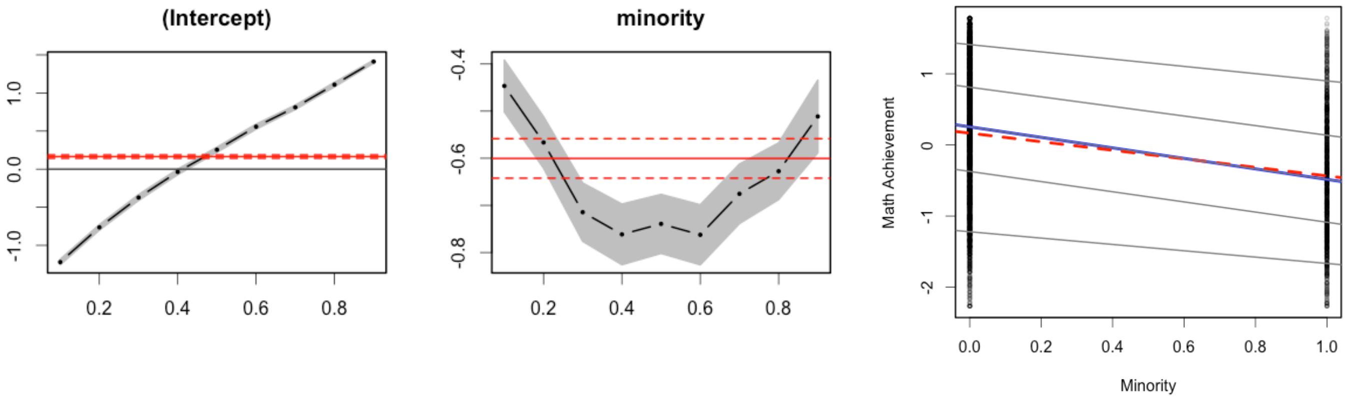
22. From these graphs, it appears as though the relationship between math achievement and SES is strongest at the median math achievement score (that's where it has the largest slope). At the median, the slope is 0.4454. At the 25th percentile (to choose another one arbitrarily), the slope is 0.4058.

    Is there a statistically significant difference between those two slopes? To check, we can compare regression models:

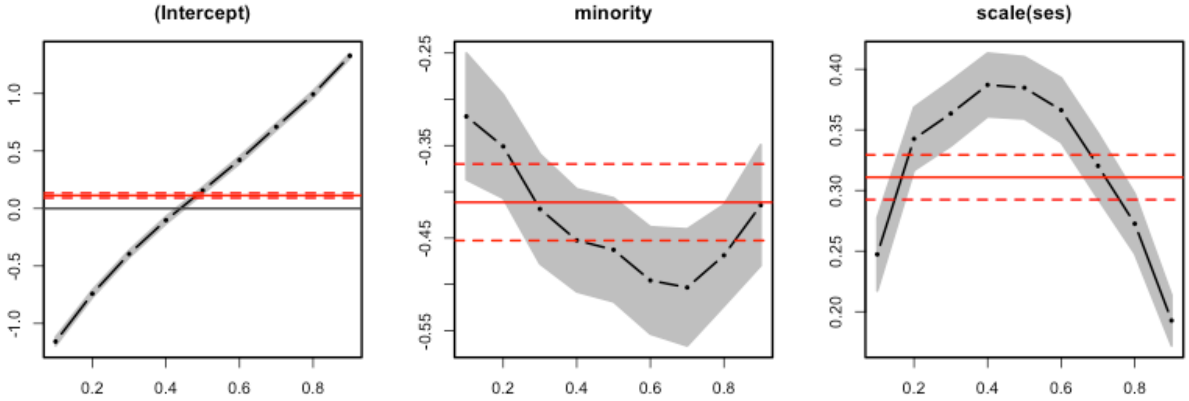
```
Quantile Regression Analysis of Deviance Table

Model: scale(mathach) ~ scale(ses)
Joint Test of Equality of Slopes: tau in {  0.25 0.5  }

  Df Resid Df F value Pr(>F)
1  1    14369    7.45 0.0064 **
```

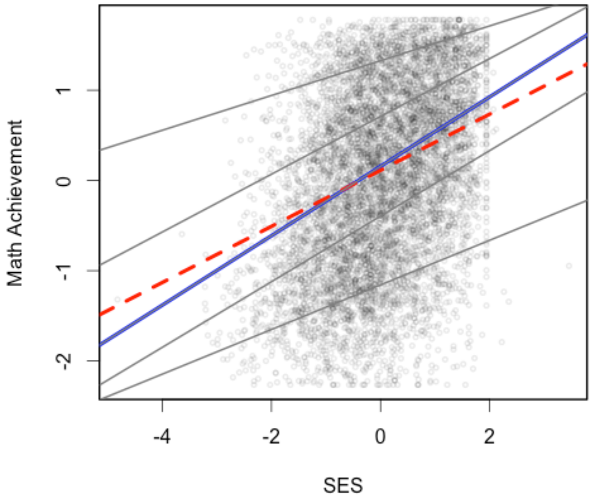
    What can we conclude from this?

23. We've investigated the math achievement gap associated with differences in socioeconomic status. Does the achievement gap for minority students also differ across levels of math achievement? Interpret the output.

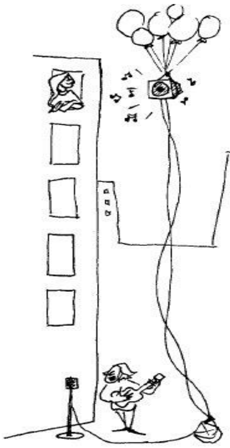


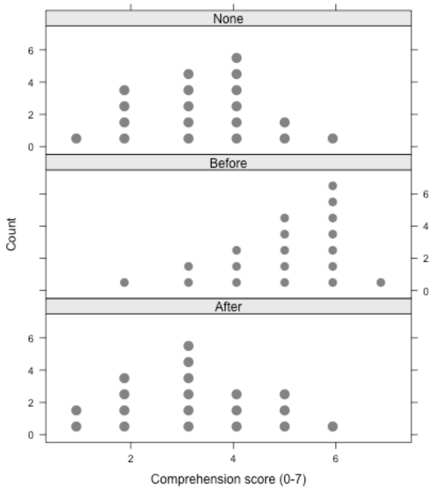
24. We can include both predictors (SES and minority) and run a multiple quantile regression analysis. Interpret.





Scenario: Recall the ambiguous prose data we used to introduce ANOVA (in activity #3):





Group	Count	Mean	Std. Dev.
None	19	3.368	1.25656
Before	19	4.947	1.31122
After	19	3.211	1.39758
Total	**57**	**M = 3.842**	**s = 1.52115**

Source	SS	df	MS	MSR (F)
Treatment	35.053	2	17.5263	10.012
Error	94.526	54	1.7505	$p = 0.0002$
Total	129.579	56	MS_{total}	$\eta^2 = 0.2705$

Data: http://www.math.hope.edu/isi/data/chap9/Comprehension.txt
Applet: http://www.rossmanchance.com/applets/AnovaShuffle.htm?hideExtras=2
Source of example: Introduction to Statistical Investigations – http://math.hope.edu/isi/
Actual Study: http://memlab.yale.edu/sites/default/files/files/1972_Bransford_Johnson_JVLVB.pdf

25. The ANOVA summary table displayed above indicates at least one of the group means differs from the others. When we conducted the ANOVA, we constructed the following model: $y_{ij} = \mu + \alpha_j + e_i$, where $\alpha_j = \mu_j - \mu$.

That model is similar to many of the linear regression models we've constructed: $y_i = b_0 + b_1 x_1 + e_i$

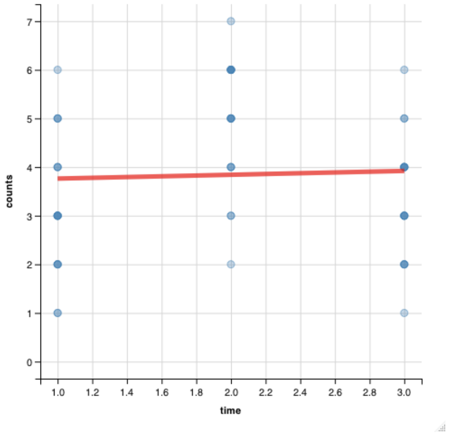
Let's take a look at the structure of our data in this example. The *original data* columns show the data that was used to generate the ANOVA summary table.

If we want to conduct a regression analysis on this data, it might make sense to convert our independent variable (condition) to numerical values. The *possible coding* column shows one way to do this. We can let 1 = after, 2 = before, and 3 = none. Since our independent variable is categorical (nominal), the actual values we use aren't meaningful (in fact, it may make more sense to code 0 = none, 1 = before, 2 = after).

	Original data		Possible coding	Dummy coding		
Subject	y = comp.score	Condition	Condition Code	After x1	Before x2	None x3
1	6	After	1	1	0	0
2	5	After	1	1	0	0
…	…	…	…	…	…	…
20	7	Before	2	0	1	0
21	5	Before	2	0	1	0
…	…	…	…	…	…	…
39	4	None	3	0	0	1
40	6	None	3	0	0	1
…	…	…	…	…	…	…

Using the possible coding listed in the table, I fit the model $y_i = b_0 + b_1(\text{condition code}) + e_i$

Here's the output. Interpret the slope coefficient and explain why this condition coding is a bad idea.



Model: comprehension = b_0 + b_1(condition code)

Formula: y = 3.68421 + 0.07895x

R^2 = 0.001828

RMSE = 1.534

F = 0.1007 (p = 0.7522)

26. When we want to include a categorical predictor in our regression model, we'll want to convert that categorical predictor to a series of dummy variables.

A dummy variable can take the value of 0 or 1 to indicate the absence or presence of a categorical effect. So, in this example, we could convert our condition variable into three dummy variables:

- Dummy variable #1 = 1 if the condition is **after** (and equals zero for the before and none categories)
- Dummy variable #2 = 1 if the condition is **before** (and equals zero for the after and none categories)
- Dummy variable #3 = 1 if the condition is **none** (and equals zero for the before and after categories)

These dummy variables are displayed in the table on the previous page.

Suppose we enter these dummy variables into our regression model: $\hat{y} = b_0 + b_1\mathrm{x}_1 + b_2\mathrm{x}_2 + b_3\mathrm{x}_3$

I entered these variables into R and obtained the following output:

```
Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.3684     0.3035  11.097 1.51e-15
x1           -0.1579     0.4293  -0.368 0.714436
x2            1.5789     0.4293   3.678 0.000542
x3                NA         NA      NA       NA
```

Why did I get an error message?

27. If we include all 3 dummy variables, we have a perfect collinearity problem. All the information we need about the condition category is contained within the first two dummy variables:

- x1 = 1 if the condition is **after** (and equals zero for the before and none categories)
- x2 = 1 if the condition is **before** (and equals zero for the after and none categories)

To demonstrate this, identify the condition (after, before, or none) for each of the following:

x1	x2	Condition
1	0	________________
0	1	________________
0	0	________________

28. Let's run our regression again, this time using only the first two dummy variables: $\hat{y} = b_0 + b_1\mathrm{x}_1 + b_2\mathrm{x}_2$

Interpret this output. What do the coefficients represent?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.3684     0.3035  11.097 1.51e-15
x1           -0.1579     0.4293  -0.368 0.714436
x2            1.5789     0.4293   3.678 0.000542
---

Residual standard error: 1.323 on 54 degrees of freedom
Multiple R-squared:  0.2705,  Adjusted R-squared:  0.2435
F-statistic: 10.01 on 2 and 54 DF,  p-value: 0.0002002
```

29. We can get an ANOVA summary table for this regression model (using the dummy variables). Compare it to the ANOVA summary table we calculated in activity #3.

Original ANOVA summary table

Source	SS	df	MS	MSR (F)
Treatment	35.053	2	17.5263	10.012
Error	94.526	54	1.7505	p = 0.0002
Total	129.579	56	MS_{total}	$\eta^2 = 0.2705$

Regression Analysis

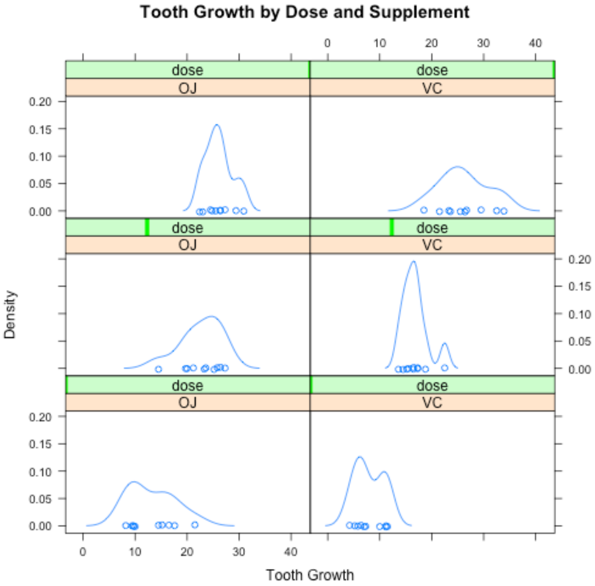
Source	SS	df	MS	MSR (F)	
model	35.053	2	17.5263	10.012	p=0.002
x1	11.369	1	11.369	6.494	p=0.014
x2	23.684	1	23.684	13.530	p=0.001
Error	94.526	54	1.7505		
Total	129.579	56	MS_{total}		

30. We can also conduct an AxB ANOVA as a regression analysis. To demonstrate, let's use the guinea pig tooth growth data we may have investigated back in the first unit.

This study investigated tooth growth in guinea pigs as a function of the type and dose of vitamin C. Guinea pigs were given a low, medium, or high dose of vitamin C either through orange juice (OJ) or a vitamin C supplement (VC).

Here's a quick summary of our data:

```
  supp dose  n  mean       sd
1   OJ  0.5 10 13.23 4.459709
2   OJ    1 10 22.70 3.910953
3   OJ    2 10 26.06 2.655058
4   VC  0.5 10  7.98 2.746634
5   VC    1 10 16.77 2.515309
6   VC    2 10 26.14 4.797731
```



We can convert our dose and supplement variables into dummy variables:

	Original data			Dummy coding		
Guinea Pig	y = tooth growth	Supplement	Dose	Supp.Code	Dose1	Dose2
1	4.2	OJ	Low	0	1	0
2	11.5	OJ	Medium	0	0	1
3	7.3	OJ	High	0	0	0
4	5.8	VC	Low	1	1	0
5	6.4	VC	Medium	1	0	1
6	10.0	VC	High	1	0	0
…	…	…	…	…	…	…

We can then fit a linear model, including the interaction terms.

$$\hat{y} = b_0 + b_1\left(\text{supp.code}\right) + b_2\left(\text{dose}_{low}\right) + b_3\left(\text{dose}_{med}\right) + b_4\left(\text{supp.code}\right)\left(\text{dose}_{low}\right) + b_5\left(\text{supp.code}\right)\left(\text{dose}_{med}\right)$$

The model results in the following coefficient estimates:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    13.230      1.148  11.521 3.60e-16
suppVC         -5.250      1.624  -3.233  0.00209
dose1           9.470      1.624   5.831 3.18e-07
dose2          12.830      1.624   7.900 1.43e-10
suppVC:dose1   -0.680      2.297  -0.296  0.76831
suppVC:dose2    5.330      2.297   2.321  0.02411
```

Interpret this output.

31. Compare the AxB ANOVA summary table with the ANOVA summary table from this regression analysis.

AxB ANOVA

Source	SS	df	MS	MSR (F)	
supplement	205.4	1	205.4	15.572	p=0.0002
dose	2426.4	2	1213.2	92.000	p<0.00001
interaction	108.3	2	54.2	4.107	0.02186
Error	712.1	54	13.2		
Total	3452.2	59	MS_{total}		

Regression

Source	SS	df	MS	MSR (F)	
supplement	205.4	1	205.4	15.572	p=0.0002
dose	2426.4	2	1213.2	92.000	p<0.00001
supp:dose	108.3	2	54.2	4.107	0.02186
Error	712.1	54	13.2		
Total	3452.2	59	MS_{total}		