CSCI 3022

intro to data science with probability & statistics

November 26, 2018

Inference & Model Selection in Multiple Linear Regression

Akaike ☺

Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Dan Larremore

Stuff & Things

1. **Homework 6** is due Friday. Final homework! :D
2. **Final Exam** (Dan’s section): December 18, Tuesday, 7:30 PM to 10 PM.
3. **Practicum** posted tonight. Due **Wednesday, 12/12, 11:55 PM.** *No late submissions accepted.*

Practicum Rules

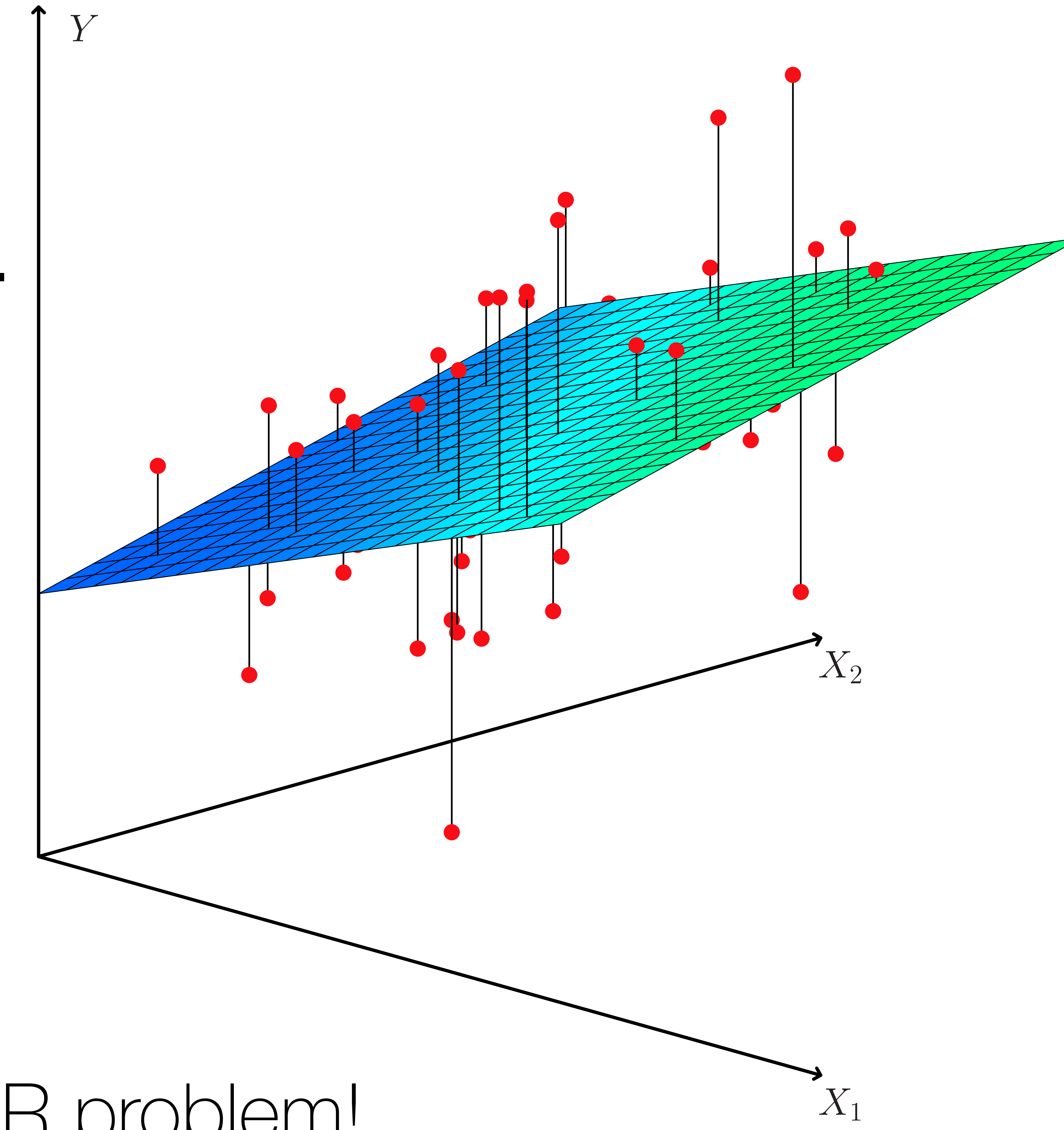
Arkaive :)

1. All work, code and analysis must be **your own**.
2. You may use your course notes, posted lecture slides, textbooks, in-class notebooks, and homework solutions as resources. You may also search online for answers to general knowledge questions like the form of a probability distribution function or how to perform a particular operation in Python/Pandas.
3. You may **not** post to message boards or other online resources asking for help.
4. **You may not collaborate with classmates or anyone else.**
5. This is meant to be like a coding portion of your final exam. So, we will be much less helpful than we typically are with homework. For example, we will not check answers, help debug your code, and so on.
6. If you have a question, send me/Tony a **private** Piazza message. If we decide that it is appropriate for the entire class, then we will add it to the **Practicum Q&A (@314)**.
7. If something is left open-ended, it is because we want to see how you approach the kinds of problems you will encounter in the wild, where it will not always be clear what sort of tests/methods should be applied. Feel free to ask clarifying questions though.

Last time on CSCI 3022:

- Multiple Linear Regression assumes that the response *y* may be affected by multiple features.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$



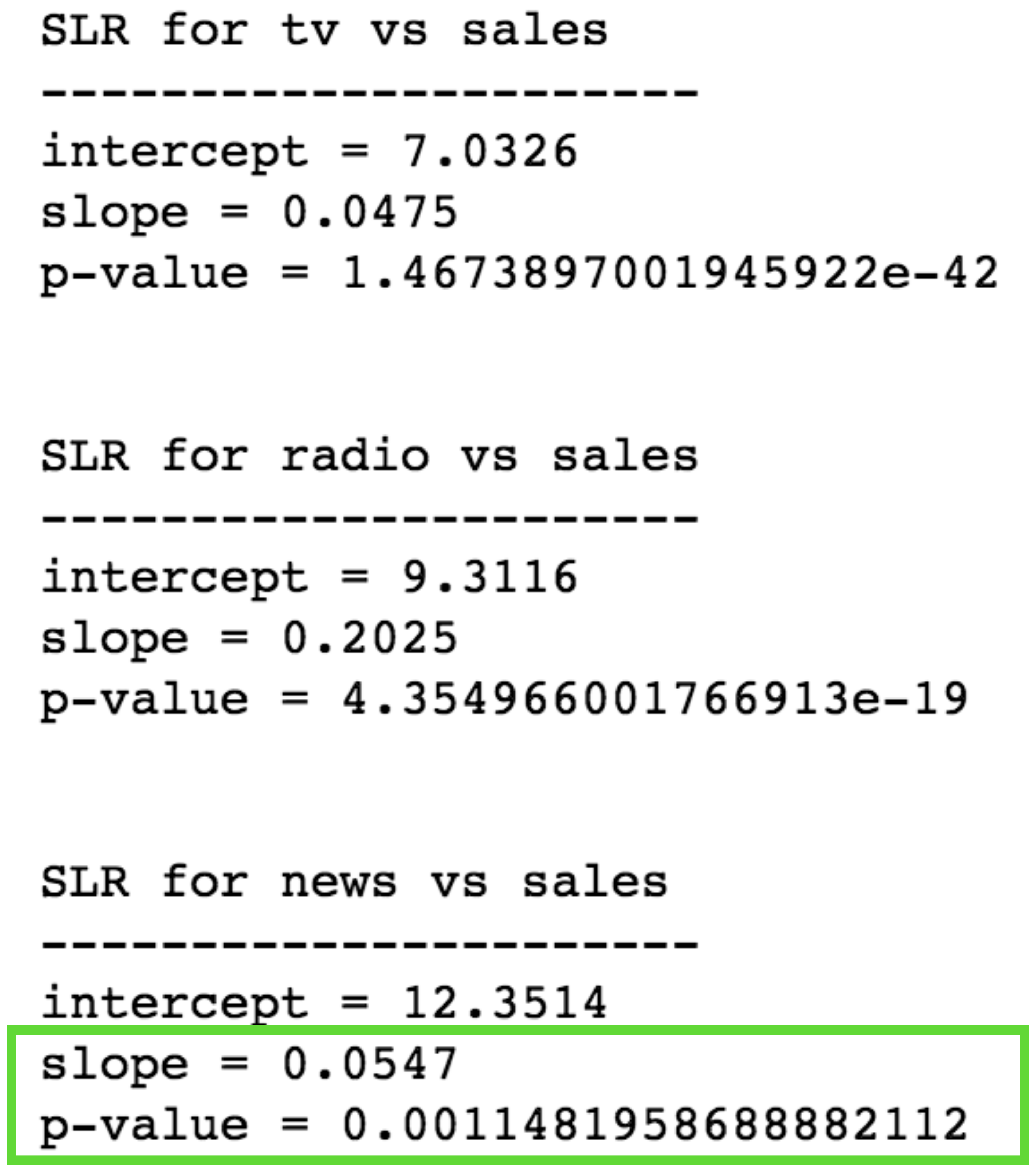
- Instead of fitting a line to the data, MLR fits a plane.

- What did we learn about MLR vs SLR?

- Note: we can cast *polynomial regression* as an MLR problem!

Recap: advertising budgets

SLR



```
SLR for tv vs sales
----------------------
intercept = 7.0326
slope = 0.0475
p-value = 1.4673897001945922e-42


SLR for radio vs sales
----------------------
intercept = 9.3116
slope = 0.2025
p-value = 4.354966001766913e-19


SLR for news vs sales
----------------------
intercept = 12.3514
slope = 0.0547
p-value = 0.001148195868882112
```

MLR

$$\texttt{sales} = 2.94 + 0.046 \times \texttt{TV} + 0.189 \times \texttt{radio} - \boxed{0.001 \times \texttt{news}}$$

Under SLR, each feature shows a significant slope.
Under MLR, the coefficient for newspapers disappears.

Covariance and Correlation of Features

- One way to discover this relationship between features is to do a **correlation analysis**. We want to know, if the value of one feature goes up is it likely that the other feature will go up as well? Similarly, we might find that if one feature goes up is it likely that the other feature will go down?

- **Def**: Let X and Y be random variables. The covariance between X and Y is given by

$$Cov(X,Y) = E\Big[(X - E[X])(Y - E[Y])\Big]$$

Let Y=X $Cov(X,X) = E\Big[(X-E[X])(X-E[X])\Big] = E\Big[(X-E[X])^2\Big] = var(X)$

- **Def**: The correlation coefficient $\rho(X,Y)$ is a measure between -1 and 1, given by

\rho

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Estimating Covariance and Correlation

- We can estimate these relationships from the data using formulas analogous to the sample variance.
- **Def**: The sample covariance is given by

$$S_{xy} = \frac{1}{n-2} \sum_{i} (x_i - \bar{x})(y_i - \bar{y})$$

- **Def**: The sample correlation coefficient is then given by

$$\hat{\rho}_{XY} = \frac{S_{xy}}{\sqrt{S_x S_y}}$$

sample variances (calculated as before)

Advertising Budget Example

- Let’s compute the pairwise correlation coefficients for the TV, radio, and newspaper spending features in the advertising data.

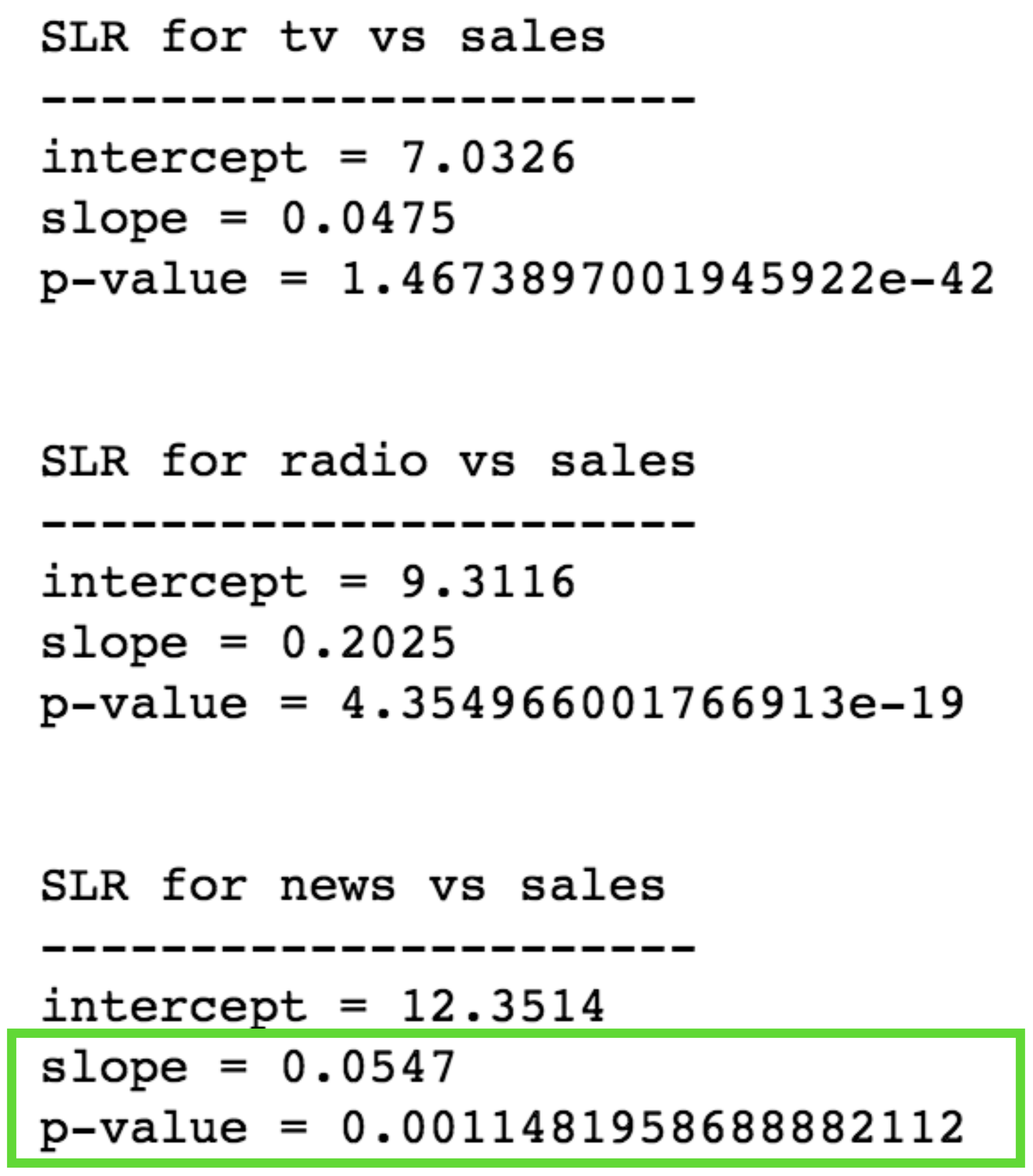
```
In [40]: dfAd[["tv", "radio", "news"]].corr()
Out[40]:
```

	tv	radio	news
tv	1.000000	0.054809	0.056648
radio	0.054809	1.000000	0.354104
news	0.056648	0.354104	1.000000

- **Question**: What do you notice? radio and news are correlated!

Recap: advertising budgets

SLR



```
SLR for tv vs sales
----------------------
intercept = 7.0326
slope = 0.0475
p-value = 1.4673897001945922e-42
```

```
SLR for radio vs sales
----------------------
intercept = 9.3116
slope = 0.2025
p-value = 4.354966001766913e-19
```

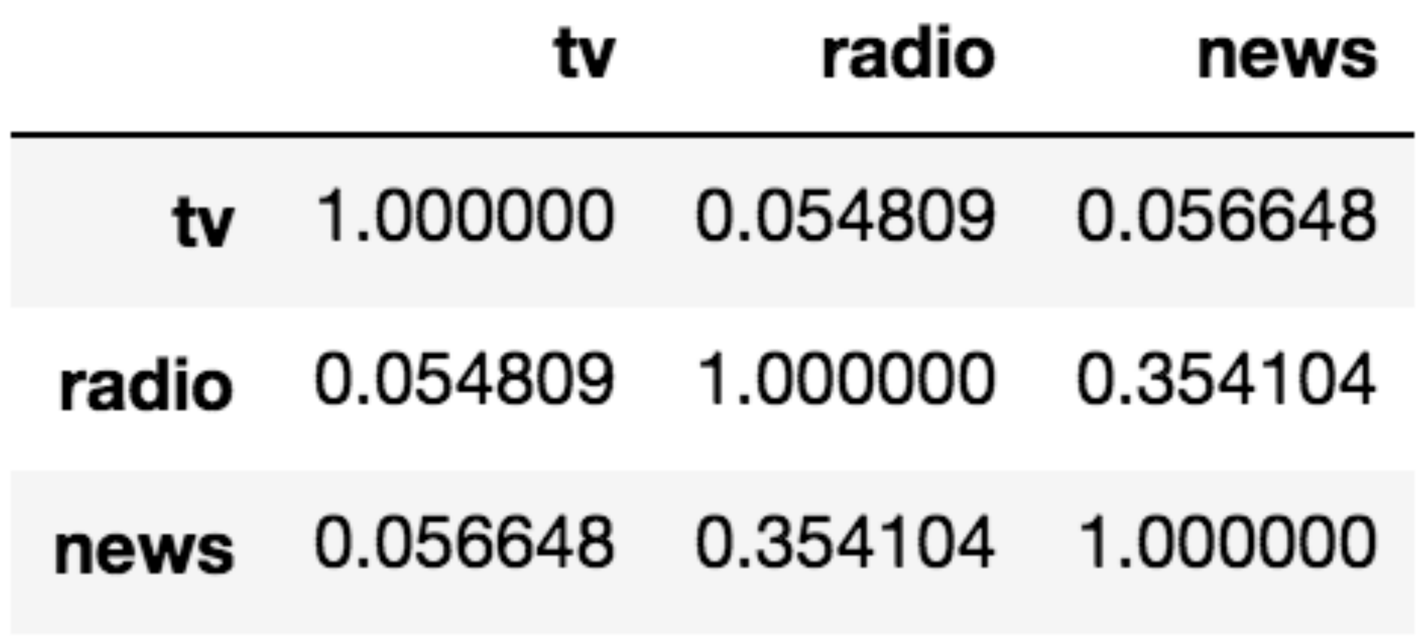
```
SLR for news vs sales
----------------------
intercept = 12.3514
slope = 0.0547
p-value = 0.0011481958688882112
```

MLR

$$\texttt{sales} = 2.94 + 0.046 \times \texttt{TV} + 0.189 \times \texttt{radio} - \boxed{0.001 \times \texttt{news}}$$

Under SLR, each feature shows a significant slope. Under MLR, the coefficient for newspapers disappears.

This is because *news* is a surrogate for *radio*, which we learned from the correlation matrix.



	tv	radio	news
tv	1.000000	0.054809	0.056648
radio	0.054809	1.000000	0.354104
news	0.056648	0.354104	1.000000

Polynomial regression

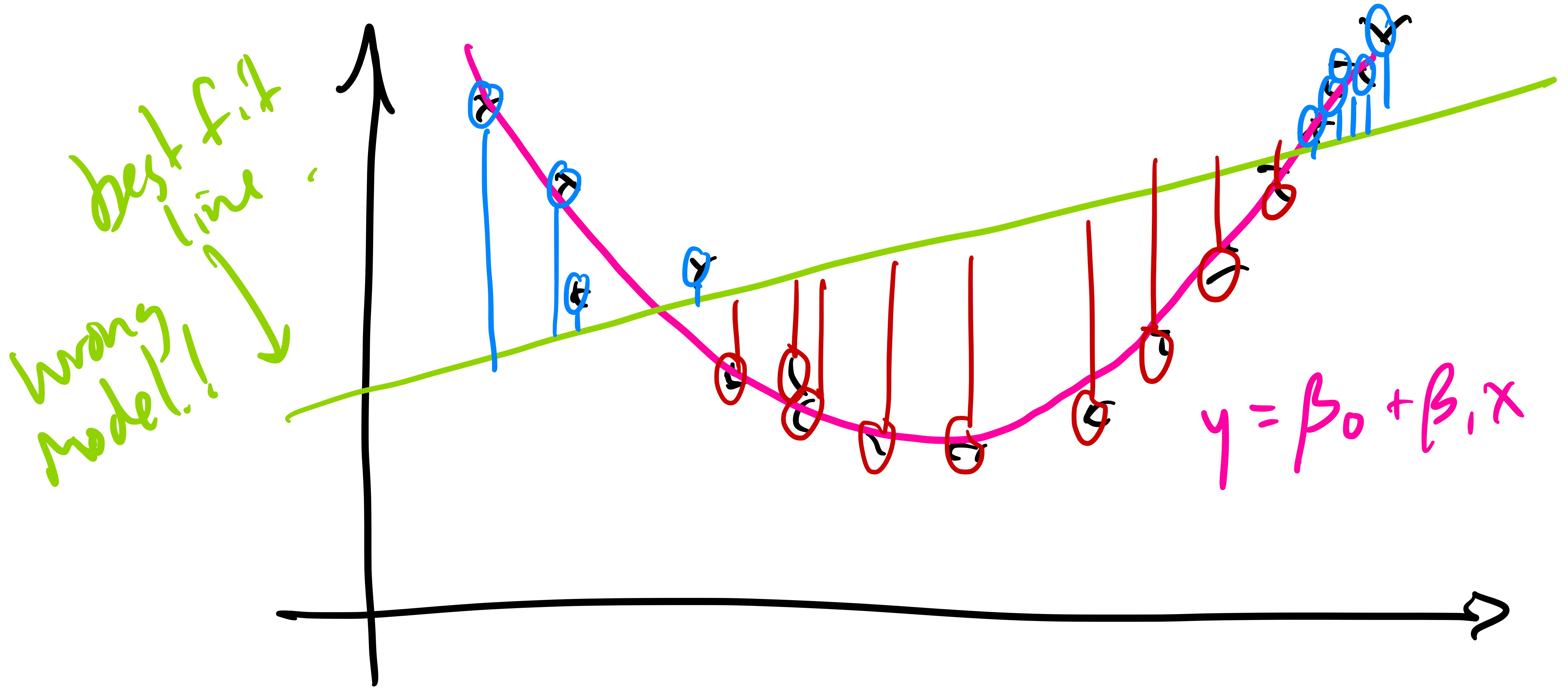
- For single-feature data, we can fit a polynomial regression model by casting it as a multiple linear regression where the additional features are powers of the original single-feature, *x*.

recall polynomial regression. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots$

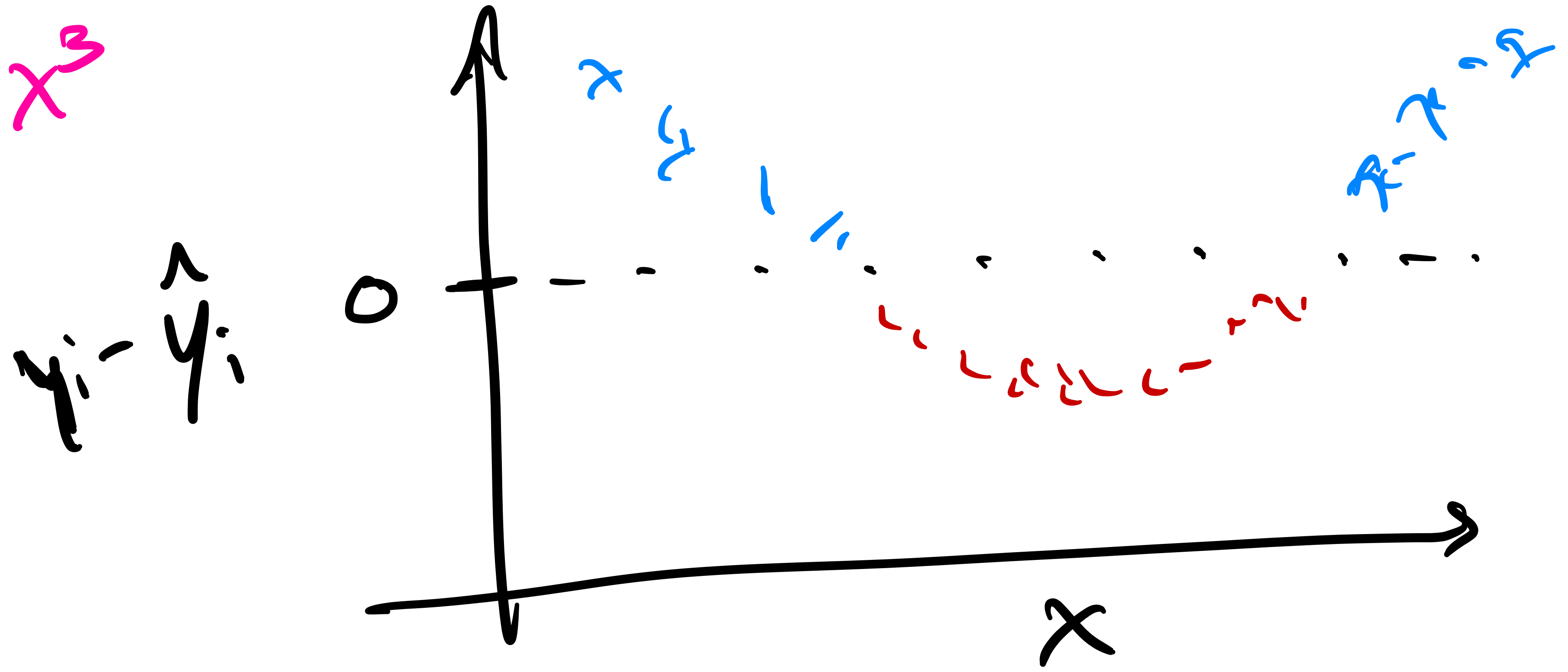
1st feature x

2nd feature x^2

3rd ~~fea~~ feature x^3



$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \underset{0}{\cancel{\beta_3}} x^3$



Using Residual Plots in Polynomial Reg.

- Recall that the assumed nature of our true model is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \dots + \beta_p x^p + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

If true model is $y = \beta_0 + \beta_1 x + \varepsilon$

and our model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

then $r = y - \hat{y} = \varepsilon \sim N(0, \sigma^2)$

If true model is $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

and our model is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

$r = y - \hat{y} \sim N(\beta_2 x^2, \sigma^2)$

See last problem on previous notebook

If I plot the residuals $r_i = y_i - \hat{y}_i$, these should be normally distributed with no dependence on x, when my model is correct.

Inference in Multiple Linear Regression

- Questions we would like to answer:
  1. Is at least one of the features useful in predicting the response?
  2. Do all of the features help to explain the response, or is it just a subset?
  3. How well does the model fit the data?

Hypothesis Testing for MLR

- Recall our question from ~~last~~ this time:

Is there a relationship between the response and predictors?

- In the simple linear regression setting, we can simply check whether $\underline{\beta_1 = 0}$.
- In the MLR setting, with *p* features (aka predictors) we n eed to ask whether *all* of the coefficients are zero:
  - H_0 : $\boxed{\beta_1 = \beta_2 = \beta_3 = \dots = \beta_p = 0}$
  - H_1 : At least one β is non-zero. $\Rightarrow$ $\boxed{\beta_j \neq 0 \text{ for at least one value of } j \neq 0}$

H_1 is $\underline{\text{not}}$ $\beta_1 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0 \dots$

Is at Least One Feature Important?

- We test the hypothesis via the F-statistic.

$$F = \frac{\dfrac{(SST - SSE)}{df_{SST} - df_{SSE}}}{\dfrac{SSE}{df_{SSE}}} = \boxed{\frac{(SST - SSE)/p}{SSE/(n-p-1)} = F}$$

- Recall:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 \qquad df_{SST} = n-1$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \dots + \hat{\beta}_p x_{ip})\right)^2$$

$$df_{SSE} = n - (p+1)$$
$$= n - p - 1$$

Is at Least One Feature Important?

- We test the hypothesis via the F-statistic.

$$F = \frac{(SST - SSE)/p}{SSE/(n-p-1)} \qquad SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$$

- Suppose H_0 were true. What would F be?

F around 1

- Suppose that H_1 were true. What would F be?

$F > 1$

The F-statistic

- We test the hypothesis via the F-statistic.

two different d.of. parameters.

$$\tilde{F} = \frac{(SST - SSE)/\boxed{p}}{SSE/\boxed{(n-p-1)}} \qquad SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad SSE = \sum_{i=1}^{n}(y_i - \hat{y})^2$$

- F distribution will give us a critical value so that we can do a p-value or rejection region test.

→ Always a one-tailed test. $F \overset{?}{\geq} F_{critical}$

compare to α like we normally would

$$1 - \text{scipy.stats.f.cdf}(\tilde{F}, p, n-p-1) \leftarrow \Pr(\tilde{F} \geq F_{p,n-p-1})$$

Is a Subset of Features Important?

- **Full Model**: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ (p=4 features in full model)
- **Reduced Model**: $y = \beta_0 + \beta_2 x_2 + \beta_4 x_4$ (k=2 features in reduced model)
- **Question**: Are the missing features important, or are we OK going with the reduced model?
- **Partial F-Test**: $H_0 : \beta_1 = \beta_3 = 0$
- Since the features in the reduced model are also in the full model, we expect the full model to perform at least as well as the reduced model.
- **Strategy**: Fit the Full and Reduced models. Determine if the difference in performance is real or due to just chance.

Is a Subset of Features Important?

- $SSE_{\text{full}} =$variation unexplained by the full model

- $SSE_{\text{red}} =$variation unexplained by the reduced model

Intuitively, if __________ is much smaller than ___________ , the full model fits the data much better than the reduced model. The appropriate test statistic should depend on the difference _______________________ in unexplained variation.

- Test Statistic: $$F = \frac{(SSE_{\text{red}} - SSE_{\text{full}})/(p-k)}{SSE_{\text{full}}/(n-p-1)} \sim F_{p-k,n-p-1}$$

- Rejection Region: $F \geq F_{\alpha,p-k,n-p-1}$

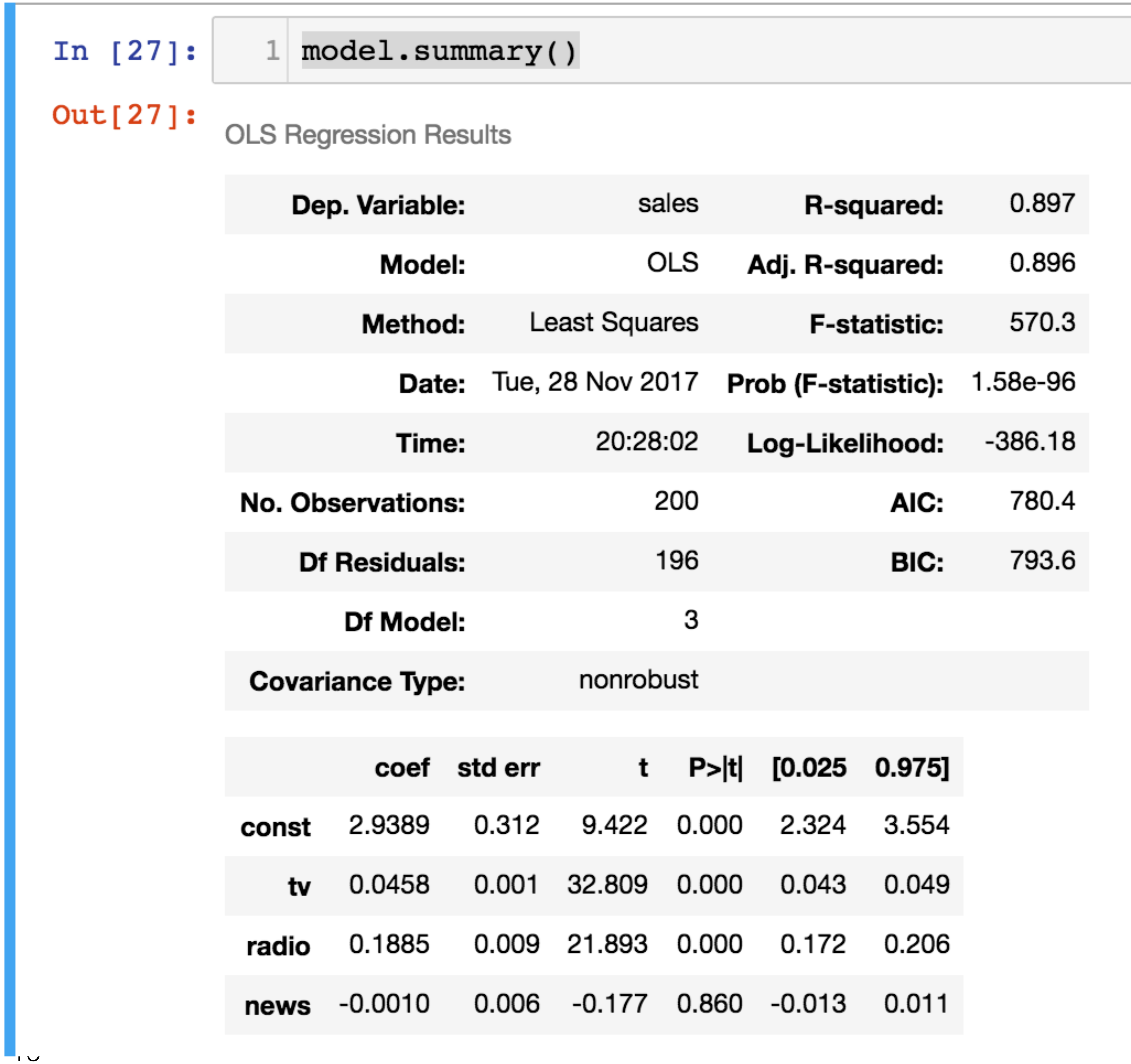
http://homepage.divms.uiowa.edu/~mbognar/applets/f.html

F… why even?

- Why compute the p-value for F-statistic when instead, we already have p-values for each of the covariates?

- Doing so would not be testing one hypothesis, but rather *p* hypotheses!

- At α=0.05, how many *p* values do we expect to be significant if the null hypothesis is, in fact, true?



```
In [27]:  model.summary()
```

```
Out[27]:
OLS Regression Results

     Dep. Variable:            sales          R-squared:     0.897
             Model:              OLS      Adj. R-squared:     0.896
            Method:    Least Squares         F-statistic:     570.3
              Date: Tue, 28 Nov 2017  Prob (F-statistic):  1.58e-96
              Time:         20:28:02      Log-Likelihood:   -386.18
  No. Observations:              200                 AIC:     780.4
      Df Residuals:              196                 BIC:     793.6
          Df Model:                3
   Covariance Type:        nonrobust

           coef  std err        t   P>|t|  [0.025   0.975]
const    2.9389    0.312    9.422   0.000   2.324    3.554
   tv    0.0458    0.001   32.809   0.000   0.043    0.049
radio    0.1885    0.009   21.893   0.000   0.172    0.206
 news   -0.0010    0.006   -0.177   0.860  -0.013    0.011
```

The road to R^2 for MLR

- Just as with simple regression, the error sum of squares is:

- It is again interpreted as a measure of how much variation in the observed y values is not explained by (not attributed to) the model relationship.
- The number of df associated with SSE is n–(p+1) because p+1 df are lost in estimating the p+1 β coefficients.

The road to R^2

- Just as before, the **total sum of squares** is:
- And the **sum of squared errors** is:
- Then the coefficient of multiple determination R^2 is:
- It is interpreted in the same way as before. (Do you remember?)

Hacking R^2

Unfortunately, there is a problem with R^2: Its value can be inflated by adding lots of predictors into the model even if most of these predictors are frivolous!

Hacking R^2

- For example, suppose y is the sale price of a house. Then:
- Sensible predictors include
  x_1 = the interior size of the house,
  x_2 = the size of the lot on which the house sits,
  x_3 = the number of bedrooms,
  x_4 = the number of bathrooms, and
  x_5 = the house's age.
- But now suppose we add in
  x_6 = the diameter of the doorknob on the coat closet,
  x_7 = the thickness of the cutting board in the kitchen,
  x_8 = the thickness of the patio slab.

Adjusted R^2

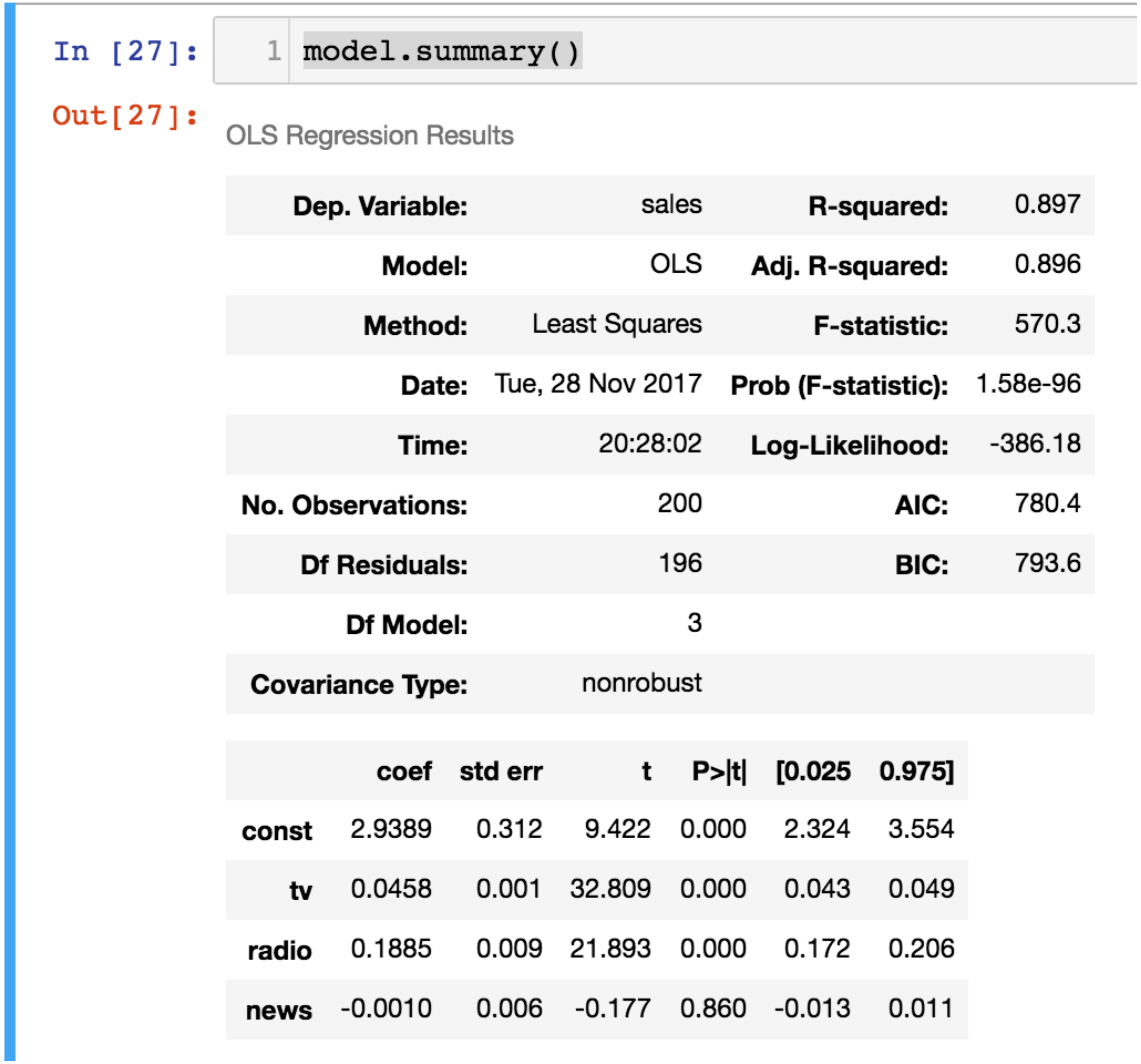
- The objective in multiple regression is not simply to explain most of the observed y variation, but to do so using a model with relatively few predictors that are easily interpreted.
- It is thus desirable to adjust R^2 to take account of the size of the model:

Adjusted R^2

- The objective in multiple regression is not simply to explain most of the observed y variation, but to do so using a model with relatively few predictors that are easily interpreted.
- It is thus desirable to adjust R^2 to take account of the size of the model:

$$R_a^2 = 1 - \frac{SSE/df_{SSE}}{SST/df_{SST}} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

Adjusted R^2



In [27]:

```
model.summary()
```

Out[27]:

OLS Regression Results

Dep. Variable:	sales	**R-squared:**	0.897
Model:	OLS	**Adj. R-squared:**	0.896
Method:	Least Squares	**F-statistic:**	570.3
Date:	Tue, 28 Nov 2017	**Prob (F-statistic):**	1.58e-96
Time:	20:28:02	**Log-Likelihood:**	-386.18
No. Observations:	200	**AIC:**	780.4
Df Residuals:	196	**BIC:**	793.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	2.9389	0.312	9.422	0.000	2.324	3.554
tv	0.0458	0.001	32.809	0.000	0.043	0.049
radio	0.1885	0.009	21.893	0.000	0.172	0.206
news	-0.0010	0.006	-0.177	0.860	-0.013	0.011

Deciding on important variables

- Suppose that we have 100 data points (n=100), but we have 200 different features (p=200). How can we learn which features are important and which are not?
- **Some options**:
  - Try all the possible combinations of features in models to see which gives the best fit.

Deciding on important variables

- Suppose that we have 100 data points (n=100), but we have 200 different features (p=200). How can we learn which features are important and which are not?
- **Some options**:
    - **Forward selection**:
        1. fit null model with an intercept but no predictors.
        2. fit p-SLRs, 1 for each feature. Choose the one that gives the lowest SSE.
        3. fit p-1 MLRs. Choose that which gives lowest SSE…
        4. repeat.

Deciding on important variables

- Suppose that we have 100 data points (n=100), but we have 200 different features (p=200). How can we learn which features are important and which are not?
- **Some options**:
  - **Backward selection**:
    1. Fit model with *all* predictors
    2. Remove the one with the largest *p*-value.
    3. Fit model with p-1 predictors.
    4. Remove the one with the largest *p-value…*

Quiz

1. **Advertising example**. I want to know if the set of {news,radio} have a slope that is significantly different from 0.

2. **Home prices example**. I have 1000 data points and 30 features. I want to learn the 10 most predictive and significant features.

3. **Home prices example**. I have 100 data points and 200 features. I want to learn the 20 most predictive features.

4. **Shark attacks example**. I have 50 shark attacks, and I have 20 features *but they are unlabeled*. I want to compute how well my model fits the data.