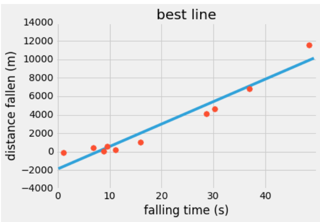
STAT/CS 94 Fall 2015 Adhikari HW12, Due: 12/02/15

Problem 1 Guessing Gravity

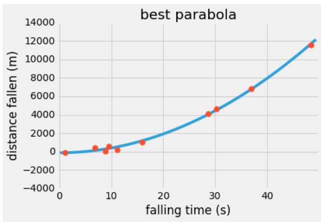
Suppose you are an early natural scientist trying to understand the relationship between the length of time (`t`, in seconds) an initially-stationary object above Earth's surface spends in free fall and the distance (`d`, in meters) it travels in that time. (Newton will later predict, using calculus and a model of physics, that the relationship is $d = \frac{1}{2}gt^2$, where g is a constant related to the gravity of Earth.) You run experiments in which you drop an iron ball 10 times from a *very* tall cliff; each time, you choose a time randomly between 0 and 50 seconds and measure the distance it has fallen at that time. Your distance measurements rely on a human assistant with a stopwatch standing on the ground, so they are somewhat noisy.

You have three hypotheses: (i) distance is a linear function of falling time; (ii) distance is a quadratic function of falling time; or (iii) distance is a 9th-degree polynomial function of falling time. To test these, you decide to find the function that fits the data most closely under each hypothesis, in the sense of minimizing the average squared residual. [1] You plot the curves and the data, getting the following three pictures:

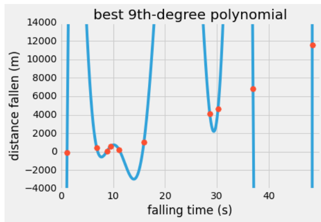
(i)



(ii)



(iii)



(The polynomial doesn't actually have discontinuities anywhere; it just varies so sharply that we couldn't fit the whole curve on the same scale as the linear and quadratic curves without making those curves look very flat.)

(a) Rank the curves by average squared residual, least to greatest. If you think there is a tie between any of the curves, say that.

(b) Informally, which hypothesis do you think is most supported by these data? Why? (You don't need to do a formal hypothesis test.)

(c) Suppose you ran another copy of the experiment, drew *the curves from the first copy of the experiment* (the ones displayed in the pictures above) over the 10 points *from the second copy of the experiment* (not pictured), and computed their residuals. Rank the curves by the average squared residual *you would expect to see*, least to greatest. If you think any of the curves have typically about the same average squared residual, say that.

[1] You don't need to know anything about polynomials or fitting them for this question, but here are some more details about what this means. For hypothesis (i), you would find the least-squares fit line as usual, with a slope and an intercept. Note that a line is the graph of a degree-1 polynomial. For hypothesis (ii), you would find the parabola that minimizes the average squared residual; a parabola is the shape of 2nd-degree polynomial curves like $d = at^2 + bt + c$, so it has 3 parameters (a, b, and c) to fit. A 9th-degree polynomial curve looks like $d = at^9 + bt^8 \cdots + ht^2 + it + j$, so when fitting the curve for hypothesis (iii) you would have 10 parameters to choose.

Problem 2 Dummy Divisor

This problem and the next two problems use the table baby, which has been analyzed several times in class. For reference, here are the top 10 rows of the table:

```
baby = Table.read_table('baby.csv')
baby
```

Out[25]:

birthwt	gest_days	mat_age	mat_ht	mat_pw	m_smoker
120	284	27	62	100	0
113	282	33	64	135	0
128	279	28	64	115	1
108	282	23	67	125	1
136	286	25	62	93	0
138	244	33	62	178	0
132	245	23	65	140	0
120	289	25	62	125	0
143	299	30	66	136	1
140	351	27	68	120	0

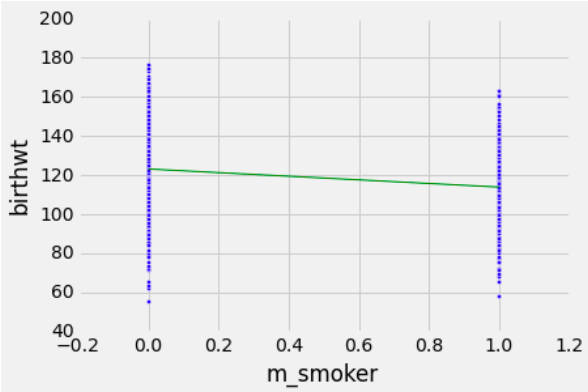
... (1164 rows omitted)

The variables are:

1. `birthwt`: baby's birthweight, in ounces
2. `gest_days`: number of gestational days
3. `mat_age`: mother's age in completed years
4. `mat_height`: mother's height in inches
5. `mat_pw`: maternal pregnancy weight in pounds
6. `m_smoker`: whether the mother is a smoker (1) or nonsmoker (0)

Now, back to this problem. From a geometric perspective it seems rather silly to perform a linear regression of `birthwt` on `m_smoker`. However, `m_smoker` has been coded as a numerical variable, so it is possible to do the regression; and indeed, the slope of the regression line has a clear interpretation. The regression has been performed below. The function `regress` returns the slope and the intercept of the regression line. Explain why the slope is the same (apart from rounding) as the output of the last line of code in the figure.

```
In [39]: scatter_fit(baby, 'm_smoker', 'birthwt')
```




```
In [40]: regress(baby, 'm_smoker', 'birthwt')

Out[40]: array([  -9.26614257,  123.08531469])

In [41]: smokers = baby.where(baby['m_smoker'],1)
         nonsmokers = baby.where(baby['m_smoker'],0)

In [43]: np.mean(smokers['birthwt']) - np.mean(nonsmokers['birthwt'])

Out[43]: -9.2661425720249184
```

Problem 3 Judging Gestation

The correlation matrix is given below.

	birthwt	gest_days	mat_age	mat_ht	mat_pw	m_smoker
birthwt	1.000000	0.407543	0.026983	0.203704	0.155923	-0.246800
gest_days	0.407543	1.000000	-0.053425	0.070470	0.023655	-0.060267
mat_age	0.026983	-0.053425	1.000000	-0.006453	0.147322	-0.067772
mat_ht	0.203704	0.070470	-0.006453	1.000000	0.435287	0.017507
mat_pw	0.155923	0.023655	0.147322	0.435287	1.000000	-0.060281
m_smoker	-0.246800	-0.060267	-0.067772	0.017507	-0.060281	1.000000

Based on this matrix, a researcher decides to regress `birthwt` on `gest_days` and `m_smoker`. The results are given below.

OLS Regression Results

Dep. Variable:	birthwt	**R-squared:**	0.216
Model:	OLS	**Adj. R-squared:**	0.214
Method:	Least Squares	**F-statistic:**	161.0
Date:	Mon, 23 Nov 2015	**Prob (F-statistic):**	1.71e-62
Time:	20:48:45	**Log-Likelihood:**	-4937.3
No. Observations:	1174	**AIC:**	9881.
Df Residuals:	1171	**BIC:**	9896.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
const	-3.1849	8.329	-0.382	0.702	-19.527 13.157
gest_days	0.4512	0.030	15.200	0.000	0.393 0.509
m_smoker	-8.3744	0.973	-8.603	0.000	-10.284 -6.464

(a) The observed slope of the variable `gest_days` is 0.4512. Assuming that the regression model holds, do the data support the null hypothesis that the true slope of this variable is 0? Justify your answer using the t-statistic as well as the confidence interval.

(b) Explain what the slope -8.3744 means for the babies of smokers and non-smokers; use the correlation matrix to support your explanation.

Problem 4 Guessing Girth

Below is the result of regressing `birthwt` on `gest_days`, `mat_ht`, and `m_smoker`.

OLS Regression Results

Dep. Variable:	birthwt	**R-squared:**	0.248
Model:	OLS	**Adj. R-squared:**	0.246
Method:	Least Squares	**F-statistic:**	128.7
Date:	Mon, 23 Nov 2015	**Prob (F-statistic):**	4.59e-72
Time:	21:34:15	**Log-Likelihood:**	-4912.4
No. Observations:	1174	**AIC:**	9833.
Df Residuals:	1170	**BIC:**	9853.
Df Model:	3		
Covariance Type:	nonrobust		

to scroll output; double click to hide

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
const	-83.0112	13.874	-5.983	0.000	-110.231 -55.791
gest_days	0.4363	0.029	14.969	0.000	0.379 0.493
mat_ht	1.3120	0.184	7.114	0.000	0.950 1.674
m_smoker	-8.5226	0.954	-8.936	0.000	-10.394 -6.651

(a) Why do you think `mat_pw` was not included in the list of predictor variables?

(b) Use this regression to predict the birth weight of a baby who has 290 gestational days and whose mother is a non-smoker 62 inches tall.

(c) Repeat part (b) assuming that all the information remains the same except that the mother is a smoker.

STAT/CS 94 Fall 2015 Adhikari HW12, Due: 12/02/15

NAME: SID:

Problem 1 Guessing Gravity

(a)

(b)

(c)

Problem 2 Dummy Divisor

Problem 3 Judging Gestation

(a)

(b)

Problem 4 Guessing Girth

(a)

(b)

(c)