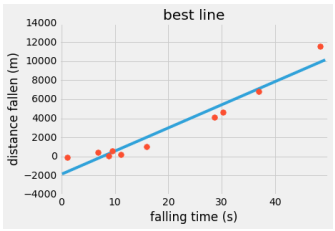
STAT/CS 94 Fall 2015 Adhikari HW12, Due: 12/02/15

Problem 1 Guessing Gravity

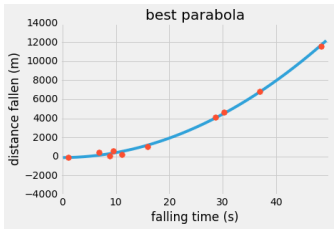
Suppose you are an early natural scientist trying to understand the relationship between the length of time (`t`, in seconds) an initially-stationary object above Earth's surface spends in free fall and the distance (`d`, in meters) it travels in that time. (Newton will later predict, using calculus 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:

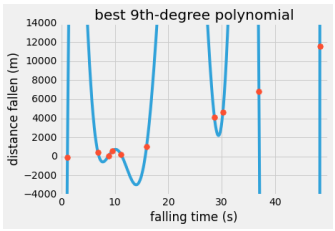
(i)



(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.

Answer:

(a) iii, ii, i.

(b) The 9th-degree polynomial does fit these data best. However, there are three reasons not to think that this is evidence for the hypothesis that the true relationship is a 9th-degree polynomial:

[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.

(i) We have a very strong physical intuition that things drop further when you let them fall longer, and the best-fit 9th-degree polynomial model strongly violates that intuition.

(ii) The problem said that the observations were a little noisy, so even if we used exactly the right physical model, we wouldn't expect it to to exhibit a perfect fit. So it's not clear that the residuals being smaller than the other curves' residuals is a good sign for this hypothesis.

(iii) A 9th-degree polynomial will fit *any* 10 data points perfectly, as long as no two horizontal-axis values are exactly the same. So the fact that the 9th-degree polynomial is a perfect fit is not really evidence for that hypothesis one way or another. (You probably learned this fact at some point in high school Algebra, but you might have forgotten it.)

By contrast, the quadratic fits pretty well, and the residuals for the line look like they have a curved pattern. So these data look like moderate evidence for the quadratic model. We would probably need more data if we wanted to check whether the 9th-degree polynomial was a better model.

The important thing to take away here is that this is an example of *overfitting*. The 9th-degree polynomial has, informally, many more degrees of freedom in fitting the available data, so it is more likely to fit them well, even if it's not the right physical model. In general, you should require more evidence before you accept a model that has more degrees of freedom.

(c) ii, i, iii. (The argument here is similar to that in (b), though we should note that there is difference between the hypotheses that "there is *some* 9th-degree polynomial that is a good fit for the underlying physical process" (the scientist's question, which is what we asked about in the previous part) and "the *particular* 9th-degree polynomial that we found using our dataset predicts future observations well" (the forecaster's question, which is what we're asking here). In any case, it's clear that the 9th-degree polynomial shoots off to very large values between, for example, 17 and 28. Visually, the data don't give strong evidence that the real relationship between time and distance looks like that in that region; they almost lie on a smooth curve. And we would expect, before looking at any data, that distance would increase smoothly with time. So if we were to run the experiment and get any times between 17 and 28, the residuals for the 9th-degree polynomial would likely be much higher. On the other hand, by our above reasoning, the quadratic curve (ii) looks like a pretty plausible fit to the data, so it would probably have smaller residuals than the line (i).)

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

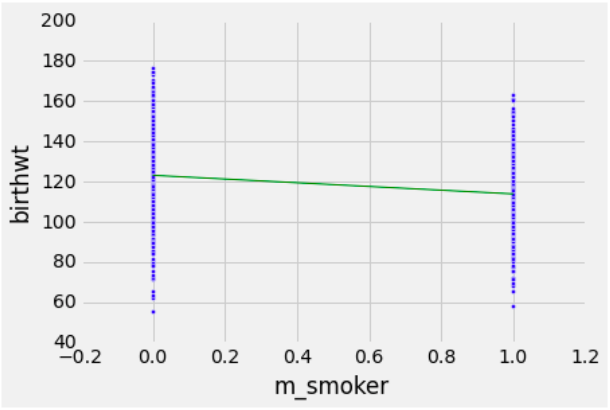
... (1164 rows omitted)

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
```

Answer: Recall our initial discussions in class about predicting points in scatter plots. If your data consists of several vertical strips, then your best prediction for each strip (specifically, the prediction that minimizes the squared residuals from that prediction among the points in that strip) is that strip's mean. (We did not prove this fact, but you can prove it if you can do calculus.) In general we can't accomplish that with a line, since we generally can't find a line that passes through 3 or more specified points. But we can always find a line that passes through *2* specified points, and in this case, we have only 2 vertical strips (at `m_smoker == 0` and `m_smoker == 1`), so the best-fit line can and therefore does pass through the means of those strips. The slope of a line that passes through the two points (x_1, y_1) and (x_2, y_2) is $\frac{y_2 - y_1}{x_2 - x_1}$, so a line that passes through $(0, \texttt{np.mean(nonsmokers['birthwt'])})$ and $(1, \texttt{np.mean(smokers['birthwt'])})$ has slope equal to the difference in those two means.

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.

Answer:

(a) The approximate 95% confidence interval given in the table is $[0.393, 0.509]$, which does not include 0. So we can reject the null hypothesis (with an approximate P-value that is less than .05). This is similar to the method discussed in the textbook [2]; the only difference is that we don't know that the confidence interval was generated with the bootstrap.

Using the t statistic gives us an hypothesis test that is closer in flavor to the ones we saw early in this class. The t statistic has (under some assumptions) a certain distribution under the null hypothesis, and

[2] See `http://data8.org/text/4_prediction.html#Is-there-a-linear-trend-at-all?`.

we calculate a P-value as the probability of getting a t statistic as extreme as or more extreme than the one we computed on the sample, if the null hypothesis were true. The table tells us this P-value directly (it's called "`P>|t|`" in the table): 0.000. (This is surely rounded down from some small positive number, and not actually 0.) So using this t-test we could reject the null hypothesis with a P-value threshold of 5%.

(b) Using multiple linear regression on our data, we predict that the babies of smokers weight 8.3744 fewer ounces than those of nonsmokers *with the same number of gestation days*. Since the correlation between `gest_days` and `m_smoker` is small, it is reasonable to talk about holding gestational days constant and varying `m_smoker`. (If they were very correlated, it might not make sense to imagine holding one constant while varying the other.)

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		

	coef	std err	t	P>|t|	[95.0% Conf. Int.]
to scroll output; double click to hide					
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.

Answer:

(a) We have already included `mat_ht` as a variable in the regression, and `mat_ht` is highly correlated with `mat_pw` (as we would expect!). We saw in lecture that *multicollinearity* (high correlation between regressor variables) can affect the accuracy of the coefficients on `mat_ht` and `mat_pw` as estimates of the real coefficients in the population from which `baby` was sampled. So if we want to accurately estimate the coefficient on `mat_pw` we would find in a regression using the entire population, it is a good idea not to include `mat_wt`.

(b) In Python: `0.4363*290 + 1.3120*62 + (-8.5226)*0 + (-83.0112)`, or 124.86 ounces (around 7.8 pounds).

(c) In Python: `0.4363*290 + 1.3120*62 + (-8.5226)*1 + (-83.0112)`, or 116.34 ounces (around 7.3 pounds).