

Statistics 3540
Assignment #2: Solutions

1. Find least squares estimate of β_1 in

$$y_i = \beta_1 x_i^2 + e_i$$

Solution: Define $SS(\beta_1) = \sum_i (y_i - \beta_1 x_i^2)^2$. Then

$$\begin{aligned}\frac{dSS(\beta_1)}{d(\beta_1)} &= -2 \sum_i x_i^2 (y_i - \beta_1 x_i^2) = 0 \\ \sum_i x_i^2 (y_i - \beta_1 x_i^2) &= 0 \\ \beta_1 \sum_i x_i^4 &= \sum_i x_i^2 y_i \\ \hat{\beta}_1 &= \frac{\sum_i x_i^2 y_i}{\sum_i x_i^4}\end{aligned}$$

Technically, at this point, we should take second derivatives to make sure that $\hat{\beta}_1$ is a global min (and not max), but I didn't require you to do this.

2. For simple linear regression, show that $\sum_i y_i = \sum_i \hat{y}_i$.

The real temptation in this question is to get to a point where you have $\sum_i e_i$, then say that $\sum_i e_i = 0$. Unfortunately, we can't say that. In class we stated that the **residuals** sum to 0: $\sum_i e_i = 0$, where $e_i = y_i - \hat{y}_i$. However, the same doesn't hold for the model errors (ϵ_i). These are random with mean 0, but we don't know that they sum to 0.

Solution: We know that $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. We will also need to use the following rules on summations:

$$\begin{aligned}\sum_{i=1}^n x &= nx, \quad \sum_i (ax_i + by_i) = a \sum_i x_i + b \sum_i y_i \\ \sum_{i=1}^n \hat{y}_i &= \sum_i (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\ &= \sum_i \hat{\beta}_0 + \hat{\beta}_1 \sum_i x_i \\ &= n\hat{\beta}_0 + \hat{\beta}_1 n\bar{x} \quad (\text{rearrange } \bar{x} = \sum_i x_i/n) \\ &= n(\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 n\bar{x} \\ &= n\bar{y} = n \sum_i y_i/n = \sum_i y_i\end{aligned}$$

3. Use data in Table 5.7, p. 264.

(a) Fit model

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + e$$

to data. Some of the Minitab output is below.

The regression equation is

Sales = 48.4 - .00341 space - 0.574 pres - 0.0509 park

Predictor	Coef	SE Coef	T	p
Constant	48.365	6.955	6.95	0.000
space	-0.003412	0.001255	-2.72	0.015
pres	-0.5743	0.1040	-5.52	0.000
park	-0.0591	0.06140	-0.83	0.419

s = 9574 R-sq = 96.9% R-sq(adj) = 96.0%

From the output, the least squares line is

$$\hat{y} = 48.4 - .00341x_1 - 0.574x_2 - 0.0509x_3$$

(b) From the output: $R^2 = 96.9\% = 0.969$. Therefore, 96.9% of the variation in prescription sales can be explained by the regression line relating floor space, prescription percentage and parking to sales.

Therefore this appears to be a very good model, in a general sense.

(c) Plot residuals vs. the \hat{y} values.

There does not appear to be a pattern in the plot. Therefore, it seems that most of our model assumptions (form of equation, errors are independent with constant variance) appear satisfied. Also, almost all of the residuals are in the $(-2\hat{\sigma}, 2\hat{\sigma})$ interval (or standardized residuals within $(-2, 2)$).

(d) QQ-plot of residuals.

Does this plot appear linear? It's not bad, except for a couple of points at the beginning, and the final value. So, it's probably reasonable to assume the errors are approximately normally distributed (which is what the regression model assumes).

4. DJIA data.

(a) Plot of data.

It appears that there is an increasing relationship between time and the DJIA yearly close. It actually appears quadratic, suggesting an appropriate model could be

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + e_t$$

One may also feel that the trend is actually increasing exponentially, so we may need to log-transform the data. This answer is also acceptable.

(b) Fit a simple linear regression model. I know this contradicts what (a) says about a quadratic model being appropriate.

Some output is deleted.

The regression equation is

DJIA = 38 + 123 Time

Predictor	Coef	SE Coef	T	p
Constant	37.7	196.7	0.19	0.850
Time	123.23	13.23	9.31	0.000

S = 477.109

Durbin-Watson statistic = 0.182874

Test to see if there is a positive linear relationship between time and DJIA close.

$H_o : \beta_1 = 0$ vs. $H_a : \beta_1 > 0$

$$\text{Test statistic: } t_{obs} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{SS_{xx}}} = \frac{123.23}{13.23} = 9.31 \text{ output}$$

P-value = $P(T \geq 9.31)$, using T-distribution with $(n - 2) = 23$ df.

The output gives $2P(T \geq |9.31|) = 2P(T \geq 9.31) \approx 0$.

Therefore $P(T \geq 9.31) \approx 0/2 = 0$.

Very strong evidence against H_o .

Therefore there appears to be a positive linear relationship between time and the DJIA closing value.

(c) Find a 95% prediction interval for the value of the DJIA in 1996. **Do this by hand**, and using Minitab. You may use the fact that $SS_{xx} = 2030$.

The actual closing value in 1996 was 6448.27. How does your interval compare to this observation?

We have $n = 25$ years of data, ending in 1994. So, we want to calculate this PI when $t = 27$.

By hand: From the least squares line:

$$\hat{y} = 38 + 123(27) = 3359$$

From Minitab: $s = \hat{\sigma} = 477.109$, and we're given $SS_{xx} = 2030$. Also, $x_p = 27$, $\bar{x} = 13$.

95% PI: $t_{0.025} =$ using $(n - 2) = 23$ df.

So our 95% PI for y in 1996 is:

$$\begin{aligned} & \hat{y} \pm t_{0.025} \hat{\sigma} \sqrt{1 + 1/n + (x_p - \bar{x})^2 / SS_{xx}} \\ &= 3359 \pm 2.069(477.109) \sqrt{1 + 1/25 + (27 - 13)^2 / 2030} \\ &= 3359 \pm 1052.38 = (2306.62, 4411.38) \end{aligned}$$

We see the actual value for 1996 (6448.27) falls outside of this interval.

This illustrates we must be careful when forecasting a future value, even with a model that appears "good". Our forecast is far off what actually happened.

From Minitab:

Predicted Values for New Observations

New

Obs	Fit	SE Fit	95% CI	95% PI
1	3365.0	208.4	(2933.9, 3796.1)	(2288.0, 4442.0)

Values of Predictors for New Observations

New

Obs	Time
1	27.0

The 95% PI for y is

$$\hat{y} \pm t_{\alpha/2} \sqrt{\hat{\sigma}^2 + s_y^2} = (2288.0, 4442.0)$$

Rounding has caused the difference between the Minitab result and the one found by hand.

(d) Plot of residuals vs. time.

There is a definite pattern in this plot: the residuals are positive for the first 7 years, then negative for a number of years, and so on. This indicates the errors are positively correlated.

(e) Test at $\alpha = 0.10$ if the errors are autocorrelated.

$H_o : \phi = 0$ (errors uncorrelated)

$H_o : \phi \neq 0$ (errors correlated)

Test statistic: $d_{obs} = 0.18$ (from Minitab output).

Table: $d_L = 1.29$, $d_U = 1.45$, using $n = 25$, $k = 1$.

Since $0.18 < 1.29$, reject H_o .

Errors appear to be autocorrelated.