Lesson 43: How do we perform inferences for regression?

Enter the following data into the calculator.

Year	Men	Women	Men-Wmn
2003	27.1	25.3	1.8
2002	26.9	25.3	1.6
2001	26.9	25.1	1.8
2000	26.8	25.1	1.7
1998	26.7	25	1.7
1997	26.8	25	1.8
1996	27.1	24.8	2.3
1995	26.9	24.5	2.4
1994	26.7	24.5	2.2
1993	26.5	24.5	2
1992	26.5	24.4	2.1
1991	26.3	24.1	2.2
1990	26.1	23.9	2.2
1989	26.2	23.8	2.4
1988	25.9	23.6	2.3
1987	25.8	23.6	2.2
1986	25.7	23.1	2.6
1985	25.5	23.3	2.2
1984	25.4	23	2.4
1983	25.4	22.8	2.6
1982	25.2	22.5	2.7
1981	24.8	22.3	2.5
1980	24.7	22	2.7
1979	24.4	22.1	2.3
1978	24.2	21.8	2.4
1977	24	21.6	2.4
1976	23.8	21.3	2.5
1975	23.5	21.1	2.4

In regression, there's a little catch. The best way to check many of the conditions is with the residuals, but we get the residuals only *after* we compute the regression. Before we compute the regression, however, we should check at least one of the conditions.

So we work in this order:

1. Make a scatterplot of the data to check the Straight Enough Condition. (If the relationship is curved, try re-expressing the data. Or stop.)
2. If the data are straight enough, fit a regression and find the residuals, e, and predicted values, $\hat{y}$.
3. Make a scatterplot of the residuals against x or the predicted values. This plot should have no pattern. Check in particular for any bend (which would suggest that the data weren't all that straight after all), for any thickening (or thinning), and, of course, for any outliers. (If there are outliers, and you can correct them or justify removing them, do so and go back to step 1, or consider performing two regressions—one with and one without the outliers.)
4. If the data are measured over time, plot the residuals against time to check for evidence of patterns that might suggest they are not independent.
5. If the scatterplots look OK, then make a histogram and Normal probability plot of the residuals to check the Nearly Normal Condition.
6. If all the conditions seem to be reasonably satisfied, go ahead with inference.

Inference for regression

Conditions:

Quantitative data condition

Straight enough condition: does the scatterplot look straight

Independence assumption: hard to check so instead check

Randomization Condition: individuals are a representative sample of the population...check for boring residuals

Does the plot thicken? condition: scatterplot of residuals against predicted values....there should be no "fan pattern" or growth or shrink pattern

Normal Population Assumption: (look at the normal probability plot) the distribution of the residuals should be normal.

H_0: β = 0. Slope is zero. There is no linear association between the two variables.

H_A: β ≠0

Find $t_{n-2} = \frac{b_1 - 0}{SE(b_1)}$

Get p-value --> conclusion

Confidence interval

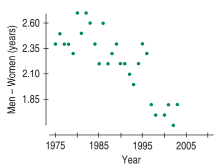
Conditions check -->Interval

$$b_1 \pm t^*_{n-2} \times SE(b_1)$$

Inference for regression is closely related to inference for means, so your understanding of means transfers directly to your understanding of regression. Here's a table that displays the similarities:

	Means	Regression Slope
Parameter	μ	β_1
Statistic	$\bar{y}$	b_1
Population spread estimate	$s_y = \sqrt{\frac{\Sigma(y - \bar{y})^2}{n - 1}}$	$s_e = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}}$
Standard error of the statistic	$SE(\bar{y}) = \frac{s_y}{\sqrt{n}}$	$SE(b_1) = \frac{s_e}{s_x\sqrt{n - 1}}$
Test statistic	$\frac{\bar{y} - \mu_0}{SE(\bar{y})} \sim t_{n-1}$	$\frac{b_1 - \beta_1}{SE(b_1)} \sim t_{n-2}$
Margin of error	$ME = t^*_{n-1} \times SE(\bar{y})$	$ME = t^*_{n-2} \times SE(b_1)$

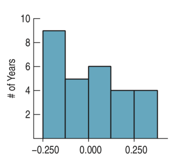
13. **Marriage age 2003.** The scatterplot suggests a decrease in the difference in ages at first marriage for men and women since 1975. We want to examine the regression to see if this decrease is significant.

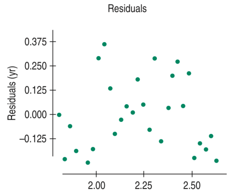


Dependent variable is: Men − Women
R squared = 65.6%
s = 0.1869 with 28 − 2 = 26 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	61.8067	8.468	7.30	≤0.0001
Year	−0.02996	0.0043	−7.04	≤0.0001

a) Write appropriate hypotheses.
b) Here are the residuals plot and a histogram of the residuals. Do you think the conditions for inference are satisfied? Explain.





c) Test the hypothesis and state your conclusion about the trend in age at first marriage.

Marriage age 2003.

a) H_0: The difference in age between men and women at first marriage has not been decreasing since 1975. $(\beta_1 = 0)$

H_A: The difference in age between men and women at first marriage has been decreasing since 1975. $(\beta_1 < 0)$

b) **Straight enough condition:** The scatterplot is not provided, but the residuals plot looks unpatterned. The scatterplot is likely to be straight enough.
Independence assumption: We are examining a relationship over time, so there is reason to be cautious, but the residuals plot shows no evidence of dependence.
Does the plot thicken? condition: The residuals plot shows no obvious trends in the spread.
Nearly Normal condition, Outlier condition: The histogram is not particularly unimodal and symmetric, but shows no obvious skewness or outliers.

c) Since conditions have been satisfied, the sampling distribution of the regression slope can be modeled by a Student's *t*-model with (28 – 2) = 26 degrees of freedom. We will use a regression slope *t*-test. The equation of the line of best fit for these data points is:

$$(\widehat{Men - Women}) = 61.8 - 0.030(Year)$$

The value of $t = -7.04$. The *P*-value of less than 0.0001 (even though this is the value for a two-tailed test, it is still very small) means that the association we see in the data is unlikely to occur by chance. We reject the null hypothesis, and conclude that there is strong evidence of a negative linear relationship between difference in age at first marriage and year. The difference in marriage age between men and women appears to be decreasing over time.