Sample Test 3 - Solution

1)

R&D Expenses

(a) The scatterplot (shown below) shows a very strong linear trend if we accept two large outliers (Microsoft and Intel).

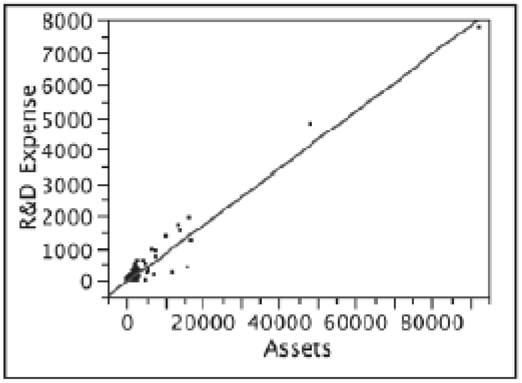
(b) The least squares equation for the shown line in the figure is
Estimated R&D Expense = 3.9819 + 0.0869 *Assets*

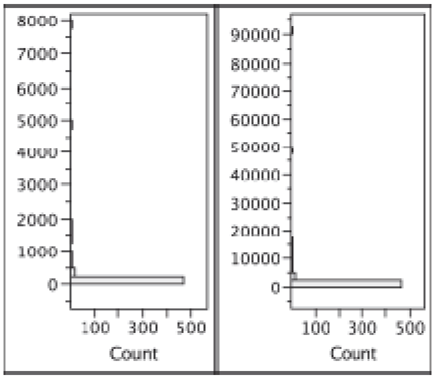
b_0 estimates that a company with no assets would still spend \$3.98 million on research and development. Although we have quite a bit of data near zero, this seems to be quite an extrapolation. Perhaps more reasonable is to interpret the intercept as a "commitment" to research that stays the same regardless of fluctuations in assets. The slope indicates that these companies on average spend about 9 cents out of each additional dollar in assets on R&D.

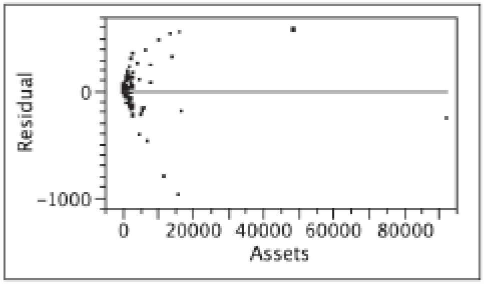
(c) $r^2 = 0.9521$ with $s_e = \$98.2915$ million. The large size of r^2 above 95% appears to reflect the fit to the large outliers rather than the variation in the smaller companies. The scatterplot "squishes" small companies into the lower left corner of the plot.

(d) Both histograms are skewed. This skewness anticipates the outlier-dominated scatterplot. Several large companies dominate both the histograms and scatterplots. (Large values in one histogram are paired with large values in the other.)

(e) The residuals show a pattern, with most of the 504 cases bunched near zero on the left. The variation appears to increase rapidly with the assets of the companies. Because the variation changes, a single summary like s_e is inadequate.







2)

Seattle homes

The reason for the transformation to cost per square feet is that the variation increases as the homes become more expensive. The lack of constant variation about the fit is more evident in the plot of the residuals from this fit.

(a) The 95% CI for fixed costs is the interval for the slope ($t_{0.025,26} = 2.056$)
57923.342 − 2.056 × 34515.8, 57923.342+ 2.056 × 34515.8 ≈ -$13,041 to $128,887

Fixed costs might be zero, but could be considerable.

(b) The confidence interval for variable costs is that for the intercept, here
155.72096 − 2.056 × 21.80695, 155.72096 + 2.056 × 21.80695≈ 111 to 201 $/SqFt

You have to pay dearly for homes in this area.

(c) Because the fixed costs might be zero, you could arguably ignore the slope in this calculation. We'll include it, taking the estimated value as our best guess. The estimated cost per square foot is then
155.72096 + 57923.342/3000 = 175.02874067 $/SqFt.

The approximate 95% prediction interval is wide, reflecting the weak fit. Because we're looking at the cost of one home, use ± 2 s_e to set the range.
175.02874067 − 2 × 41.27091, 175.02874067 + 2 × 41.27091
= 92.48692067, 257.57056067 $/SqFt ≈ 90 to 260 $/SqFt.

(d) Multiply the prior endpoints by 3000
92.48692067 × 3000 , 257.57056067 × 3000 ≈ $280,000 to $770,000

r^2	0.097731
s_e	41.27091
n	28

Term	Estimate	Std Error	t Stat	*p*-value
Intercept	155.72096	21.80695	7.14	<.0001
1/Sq Ft	57923.342	34515.8	1.68	0.1053

3)

(a) Estimated Salary = b_0 + 5 Age + 2 Test Score

(b) The indirect effect is 10 $M/Point = 2 years/point × 5 $M/year, larger than the direct effect.

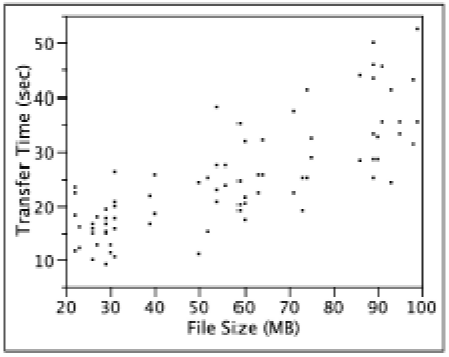
(c) The marginal effect is the direct plus indirect effect, or 10 + 2 = 12 $M/point.

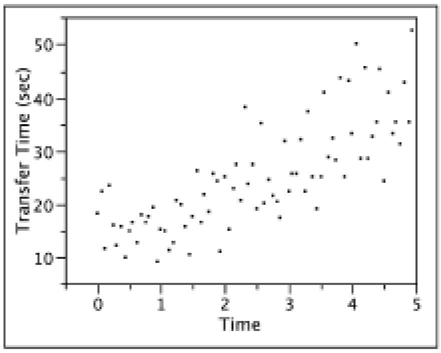
(d) You're not going to be much older, so we need the partial effect. Raising the test score by 5 points nets $10,000 annually. It's probably worth it if you're going to stay with the company long enough to earn it back.

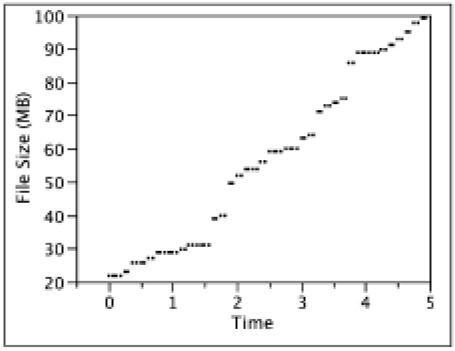
4)

Download

(a) The file sizes increased steadily over the day, meaning that these two explanatory variables are closely associated. The scatterplots of transfer time on file size and time of day seem reasonably linear, though there may be some bending in the plot of transfer time on the time of day.







(b) The marginal and partial slopes for the file size will be very different. We will not easily be able to separate their influence from one another. The file size and time of day are virtually redundant, so the indirect effect of file size will be very large.

(c) The multiple regression is

R^2	0.624569
s_e	6.283617

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	7.1388209	2.885703	2.47	0.0156
File Size (MB)	0.3237435	0.179818	1.80	0.0757
Time (hours since 8 am)	-0.185726	3.16189	-0.06	0.9533

(d) Somewhat, but not completely. The residual plot suggests slightly more variation for larger file sizes. The effect is fairly subtle and is also evident in a time plot of the residuals. There is also a slight negative dependence over time, with the residuals oscillating back in forth from positive to negative. Again, the effect is not too strong (albeit significant by the Durbin-Watson test, $D = 2.67$). The residuals appear nearly normal with no evidence of bending patterns.

(e) No. The overall F-statistic is approximately $F = (0.624/(1 - 0.624)) \times (77/2) \approx 64$ and is very significant (being much larger than 4). On the other hand, the t-statistics as seen in the tabular summary are both less than 2. Thus, we can reject H_0: $\beta_1 = \beta_2 = 0$, but cannot reject either H_0: $\beta_1 = 0$ or H_0: $\beta_2 = 0$.

(f) The key difference is the increase in the s_e of the slope. The confidence interval for the partial slope for file size from the multiple regression is $0.3237435 - 2 \times 0.179818$ to $0.3237435 + 2 \times 0.179818$, or about -.04 to 0.68 seconds per MB - a huge range that includes zero. The marginal slope is $0.3133 - 2 \times 0.0275$ to $0.3133 + 2 \times 0.0275$, or about .2583 to .3683 seconds per MB. The estimates (slopes) are about the same, but the range in the multiple regression is much larger.

(g) The direct effect of file size (from the multiple regression) is indirect effect of file size is 0.32 sec/MB. The indirect effect (from the simple regressions) is

(0.0562 hours since 8am/MB) × (-0.186 sec/hour after 8am) = -.0104532 sec/MB

is very small. The path diagram only tells you about the difference between the indirect and direct effect (slope in the simple and multiple regression), not the change in the standard errors.