

Homework 1 Solutions

January 18, 2012

Contents

1	Normal Probability Calculations	2
2	Stereo System (SLR)	2
3	Match Histograms	3
4	Match Scatter Plots	4
5	Housing (SLR)	4
6	Shock Absorber (SLR)	5
7	Participation Rate (SLR)	6
8	Height (SLR)	6
9	Market Model Example	7
10	Sample Size and SLR	8
11	T-Statistics	10
12	Housing (MLR)	10
13	Profit, Risk, and RD (MLR)	10
14	Zagat (MLR)	11
15	Salary (MLR)	13
16	Baseball (Dummy Variables)	14
17	Corn Yields (Dummy Variables)	15
18	Mid-City Housing Data (MLR, Dummy Variables)	16

1 Normal Probability Calculations

There are two ways to approach these problems: 1) pull out a calculator or 2) draw a picture and use what you know about normal distributions. A few helpful approximations: if $X \sim N(\mu, \sigma^2)$ then $P(\mu - \sigma < X < \mu + \sigma) = 68\%$, $P(\mu - 2\sigma < X < \mu + 2\sigma) = 95\%$ and $P(\mu - 3\sigma < X < \mu + 3\sigma) = 99\%$.

- (i) The normal distribution is symmetric so half its mass is above the mean. In other words, $P(X > 5) = 50\%$.
- (ii) Since $P(\mu - 2\sigma < X < \mu + 2\sigma) = 95\%$ we know that $P(X < \mu - 2\sigma \text{ or } X > \mu + 2\sigma) = 5\%$. Since the normal distribution is symmetric the left and right tails have the same mass. Thus we can conclude that $P(X > \mu + 2\sigma) = 2.5\%$.
- (iii) Using R, $P(X < 4) = \text{pnorm}(4, 5, \text{sqrt}(10)) \simeq 0.3759$.
- (iv) $P(6 < X < 15) = P(X < 15) - P(X < 6)$. Using R, $\text{pnorm}(15, 5, \text{sqrt}(10)) - \text{pnorm}(6, 5, \text{sqrt}(10)) \simeq 0.3751$.
- (v) To “standardize” a normal random variable you subtract its mean and then divide by its standard deviation, that is $Z = (X - \mu)/\sigma$. In this problem we want to transform the interval $-2 < X < 6$ to $a < Z < b$ where Z is the standardized version of X . In particular,

$$\begin{aligned} -2 &< X < 6 \\ -2 - 5 &< X - 5 < 6 - 5 \\ -7 &< X - 5 < 1 \\ -7/\sqrt{10} &< \frac{X - 5}{\sqrt{10}} < 1/\sqrt{10} \\ -7/\sqrt{10} &< Z < 1/\sqrt{10}. \end{aligned}$$

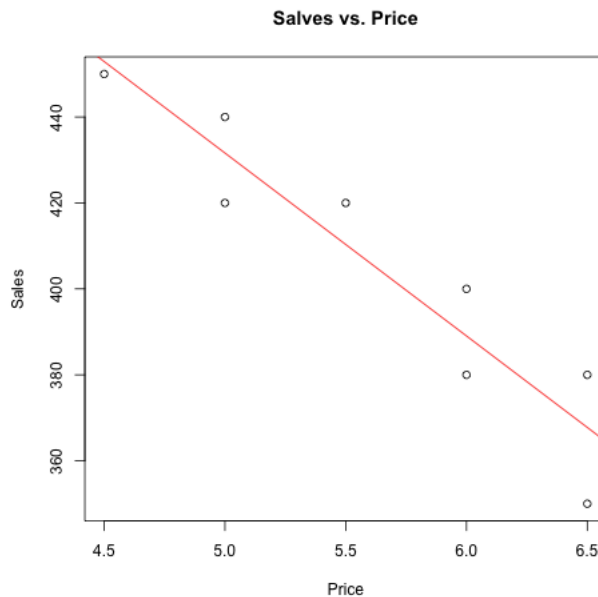
So $a = -7/\sqrt{10}$ and $b = 1/\sqrt{10}$.

2 Stereo System (SLR)

The regression is

$$\text{Sales} = \beta_0 + \beta_1 \text{Price} + \epsilon.$$

- (i) The estimates for the slope and intercept are -42.58 and 644.5 respectively.
- (ii) See the plot below.



- (iii) The 95% confidence interval for the slope is $[-58, -26]$ so it is reasonable to conclude that there is a negative relationship. This tells us that the company sells fewer stereos when the price is higher.

3 Match Histograms

These distributions look roughly normal. We can use the fact that roughly 95% of the mass of the distribution is in $[ave - 2s, ave + 2s]$ to get the average and sample variance.

In particular, if the 95% interval is $[L, R]$ then $L = ave - 2s$ and $R = ave + 2s$. Doing a bit of algebra you can derive that $(L + R)/2 = ave$ and $(R - L)/4 = s$.

X1 The 95% interval is roughly $[0, 60]$. Thus $ave \simeq 30$ and $s \simeq 15 \implies s^2 \simeq 225$.

X2 The 95% interval is roughly $[1.436, 1.445]$. Thus $ave \simeq 1.4405$ and $s \simeq 0.00225 \implies s^2 \simeq 5e - 6$.

X3 The 95% interval is roughly $[-5, 20]$. Thus $ave \simeq 7.5$ and $s \simeq 6.25 \implies s^2 \simeq 40$.

X4 The 95% interval is roughly $[80, 120]$. Thus $ave \simeq 100$ and $s \simeq 10 \implies s^2 \simeq 100$.

4 Match Scatter Plots

Looking at the slope and the intercept it appears that

- The green and blue plot will be A or D
- The black and red plots will be B or C

The red plot appears to have a larger regression standard error, which means that red corresponds to C and black corresponds to B. The R^2 seems to fit that description as well.

The regression standard errors aren't too different between the green and blue plots, but it appears that a line does a better job fitting the blue data. This suggests the blue plot should have a higher R^2 than the green plot. Hence the blue plot corresponds to D and the green plot corresponds to A.

5 Housing (SLR)

(a) $P(1.6) = 20 + 50(1.6) + \epsilon = 100 + \epsilon$.

Thus $P(1.6) \sim N(100, 15^2)$.

The 95% predictive interval is $100 \pm 2 * 15 = [70, 130]$.

(b) $P(2.2) = 20 + 50(2.2) + \epsilon = 130 + \epsilon$.

Thus $P(2.2) \sim N(130, 15^2)$.

The 95% predictive interval is $130 \pm 2 * 15 = [100, 160]$.

(c) $P [1000 \text{ Dol}] = 20 [1000 \text{ Dol}] + 50 [1000 \text{ Dol} / 1000 \text{ sq. ft.}] s [1000 \text{ sq. ft.}] + \epsilon [1000 \text{ Dol}]$.

So the slop is measured in $[1000 \text{ Dol} / 1000 \text{ sq. ft.}] = [\text{Dol} / \text{sq. ft.}]$.

(d) The intercept is measured in $[1000 \text{ Dol}]$.

(e) The units of the error standard deviation are $[1000 \text{ Dol}]$ since the standard deviation has the same units as random variable.

(f) The intercept would change from $20 [1000 \text{ Dol}]$ to $20000 [\text{Dol}]$.

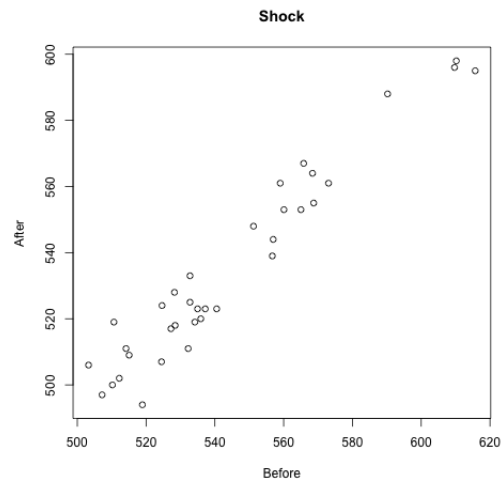
The slop would stay the same.

The error standard deviation would change from $15 [1000 \text{ Dol}]$ to $15000 [\text{Dol}]$.

(g) $P(1.6) \sim N(100, 15^2)$.

6 Shock Absorber (SLR)

- (a) It does appear that there is a linear relationship between the before and after measurements so a linear regression is appropriate. It makes sense to choose rebound, the after measurement, as the dependent variable because we want to show that you can accurately predict this value given rebound, the before measurement.



- (b) Using R we see that the confidence intervals are

- The intercept: $[-30.4, 66.8]$
- The slope: $[0.86, 1.04]$.

- (c) To test the hypothesis that $\beta_0 = 0$ we check if 0 is in the 95% confidence interval for the intercept. This is the case so we cannot reasonably reject that $\beta_0 = 0$.
- (d) The shock maker, who does not want to take the after measurement, would like the intercept to be zero and the slope to be one, in which case the before and after measurements are identical.

We saw above that we cannot reject that $\beta_0 = 0$.

To test the hypothesis that $\beta_1 = 1$ we check if 1 is in the 95% confidence interval for the slope, which it is, so we cannot reasonably reject that $\beta_1 = 1$.

- (e) The predictive estimate is $AFTER = b_0 + b_1 BEFORE$. The plug-in predictive interval is $AFTER \pm 2s$, where s is the regression standard error. According to R, the $s = 7.67$. Using R, our plug-in predictive interval is $[525, 555]$. Since this interval lies within the acceptable range for the manufacturer we can accept this shock absorber.

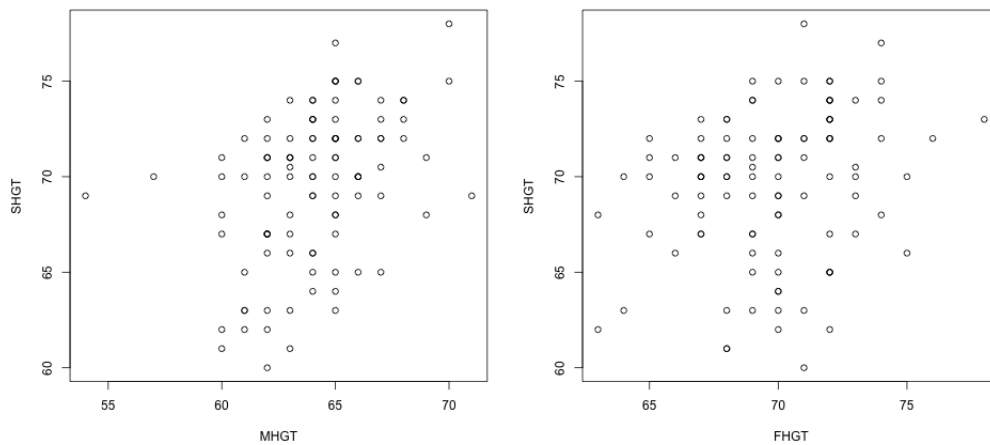
7 Participation Rate (SLR)

- (a) We know that $Corr(Y, X) = R$. From the table $R^2 = 0.397 \implies R = 0.63$.
- (b) According to the table, our prediction will be $Y = 0.203 + 0.656X$. If $X = 0.45$, then $Y = 0.498$. So the estimated participation rate in 1972 is 49.8%.
- (c) The 95% confidence interval for the slope is $b_1 \pm s_{b_1}$. Here $b_1 = 0.656$ and $s_{b_1} = 0.1961$, which gives us an interval of $[0.264, 1.05]$.
- (d) To test the hypothesis that $\beta_1 = 1$ we check if 1 is in the 95% confidence interval. It is; so we cannot reject the hypothesis that $\beta_1 = 1$.
- (e) We would expect it to be 0.397.

The R^2 in a SLR is the correlation between the independent and dependent variable, squared. But the correlation of X and Y is the same as the correlation of Y and X ; the order of X and Y do not matter. Thus it does not matter whether you are regressing X on Y or Y on X ; you get the same correlation and hence the same R^2 .

8 Height (SLR)

- (a) The mean of the mother's height is less than the mean of the father's heights.



- (b) Using R we found

- $Corr(\text{Student}, \text{Mother}) = 0.40$

- $\text{Corr}(\text{Student}, \text{Father}) = 0.23$.

As a crude estimate, this suggests that a mother's height is a better predictor a student's height.

(c) Using R we found that the means and least squares lines are

- $\text{mean}(\text{SHGT}) = 69.6$ in
- $\text{mean}(\text{MHGT}) = 64.0$ in
- $\text{mean}(\text{FHGT}) = 69.7$ in
- Student vs. Mother: $\text{SHGT} = 32.8 + 0.574 \text{MHGT}$
- Student vs. Father: $\text{SHGT} = 47.6 + 0.315 \text{FHGT}$

(d) Using R we can calculate the variance decomposition.

	SST	SSR	SSE	R^2
Mother	1536.8	240.0	1296.8	0.156
Father	1536.8	82.3	1454.5	0.054

(e) The residuals for my own height were $MRES = 0.742$ " using the Mother regression and $FRES = 1.09$ " using the Father regression.

9 Market Model Example

I chose GE, JPM, and XON as my three companies.

Name	α	95% Interval		β	95% Interval	
GE	0.008307	-0.001129577	0.01774360	1.093484	0.740402194	1.44656504
JPM	0.001065	-0.01112595	0.01325646	0.940493	0.48434462	1.39664232
XON	0.004013	-0.0039497	0.01197598	0.827766	0.5298264	1.12570506

(a) All three slope estimates are positive and close to one. The slope tells us how sensitive the returns of a stock are to the returns in the market. In particular, for every unit change in the market's returns we expect a change of β in the stock's returns. It appears that GE is most sensitive to the market and XON is least sensitive, though it is difficult to be sure about this since their 95% confidence intervals overlap.

(b) We can rewrite the model as

$$R^g = \beta R^m + N(\alpha, \sigma_g^2).$$

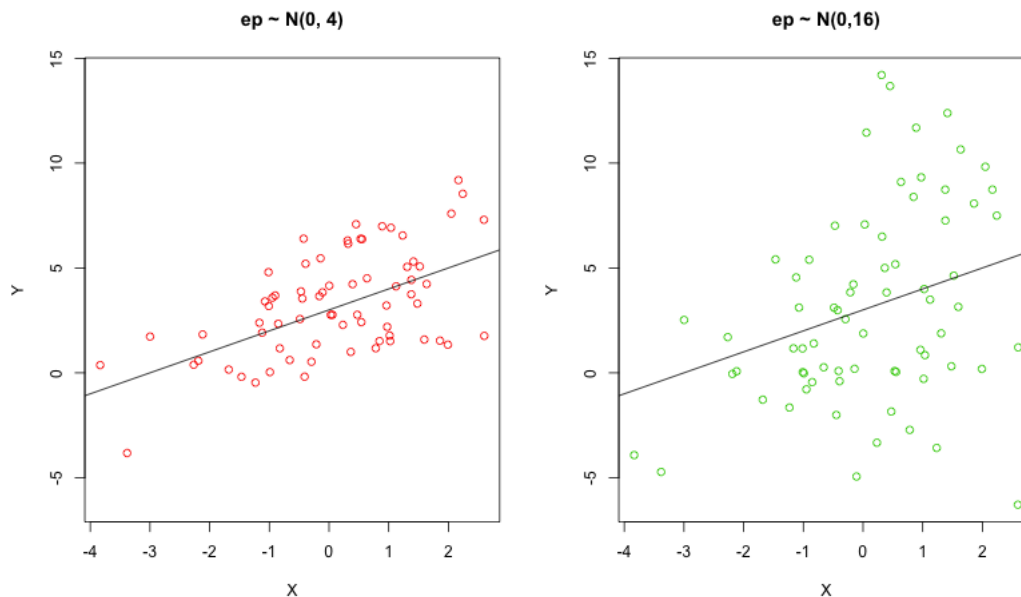
In this way we can think of the returns on a company being a combination of a market component and an idiosyncratic component. The idiosyncratic component is independent from how the market behaves. Under this interpretation α tells us about the idiosyncratic component of the mean return.

It is hard to gather any useful information from the α 's. All three confidence intervals contain zero. Further, all three confidence intervals overlap. This means that it may be difficult to identify which stock, GE, JPM, or XON, has larger idiosyncratic returns as well as whether the idiosyncratic returns are positive or negative.

- (c) The 95% confidence intervals for β are listed above.
- (d) The 99% confidence interval for the intercept in the GE regression is (0.623708460, 1.56325877). Since 1.0 is in this interval we cannot reject that $\beta = 1$. If $\beta = 1$ then a unit increase in returns on the market results in a unit increase in returns on GE.

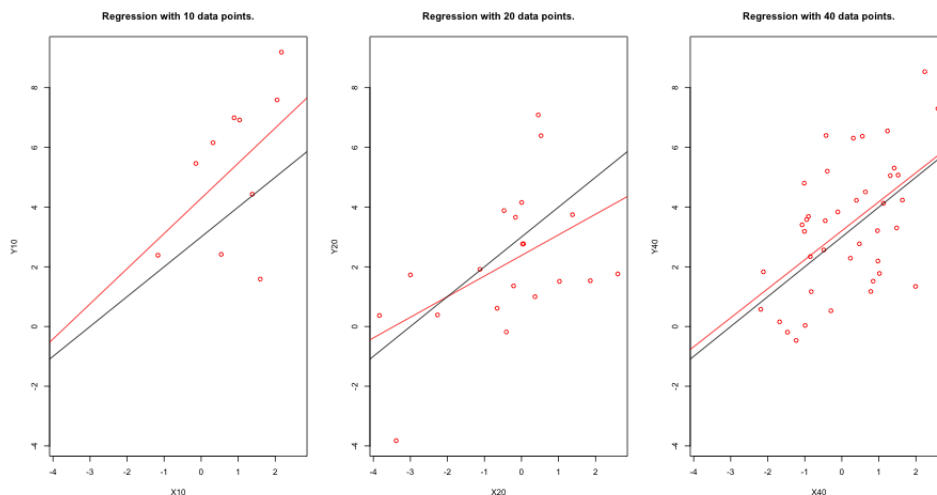
10 Sample Size and SLR

- (c), (e), and (f) : The true line is the same in both plots. However, there is less variation in Y values when the variance of the error, ϵ , is smaller.



- (g) : The estimated regression line with the most data points is the closest to the true line. This will generally be the case, though sometimes one may see that the regression using

10 or 20 data points is closest. We can see this numerically by looking at the t-stats or confidence intervals for the slope and intercept of each regression.



```
lm(formula = Y10 ~ X10)

Residuals:
    Min     1Q   Median     3Q      Max
-4.585 -1.246  1.112  1.468  2.344

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.2913    1.0024   4.281  0.00268 **
X10          1.1778    0.7689   1.532  0.16413
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 2.367 on 8 degrees of freedom
Multiple R-squared:  0.2268,    Adjusted R-squared:  0.1301
F-statistic: 2.346 on 1 and 8 DF,  p-value: 0.1641
lm(formula = Y40 ~ X40)
```

```
Residuals:
    Min     1Q   Median     3Q      Max
-3.7982 -1.3613  0.1761  1.2563  3.6054

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.2067    0.2991  10.719 4.78e-13 ***
X40          0.9716    0.2466   3.939 0.000338 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 1.88 on 38 degrees of freedom
Multiple R-squared:  0.29,    Adjusted R-squared:  0.2713
F-statistic: 15.52 on 1 and 38 DF,  p-value: 0.000338
```

```
lm(formula = Y20 ~ X20)

Residuals:
    Min     1Q   Median     3Q      Max
-3.8687 -1.5903  0.3284  1.3971  4.3931

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.3801    0.4896   4.861 0.000126 ***
X20          0.6904    0.2921   2.364 0.029541 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

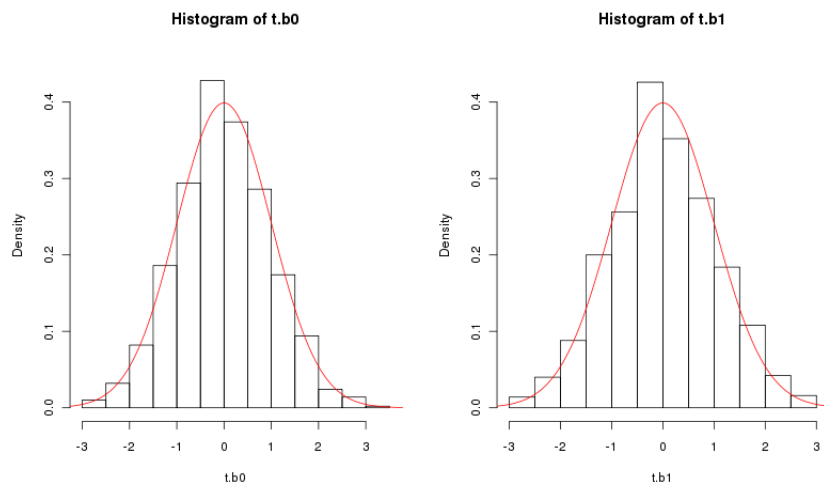
Residual standard error: 2.138 on 18 degrees of freedom
Multiple R-squared:  0.2369,    Adjusted R-squared:  0.1945
F-statistic: 5.587 on 1 and 18 DF,  p-value: 0.02954
```

(h) : Using the true model, the conditional mean give $X = -1$ is $3 - 1 = 2$. We know the standard deviation is 2. Thus our intervals are

- 68%: $2 \pm 1 * 2 = [0, 4]$.
- 95%: $2 \pm 2 * 2 = [-2, 6]$.
- 75%: $2 \pm 1.15 * 2 = [-0.3, 4.3]$.

11 T-Statistics

The histograms we produce do look like standard normal random variables.



12 Housing (MLR)

(a) $20 + 50 * 1.6 + 10 * 3 + 15 * 2 = 160$. So

Price given size = 1.6, nbed = 3, nbath = 2 is $N(160, 10^2)$.

(b) The 95% predictive interval is $160 \pm 2 * 10 = [140, 180]$.

(c) The conditional mean is $20 + 50 * 2.6 + 10 * 4 + 15 * 3 = 235$.

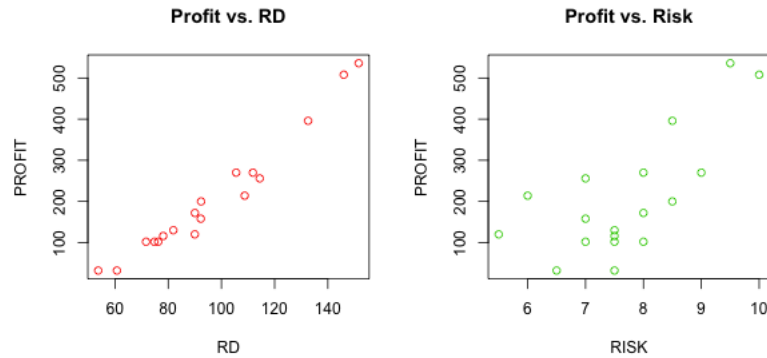
The 95% predictive interval is $235 \pm 2 * 10 = [215, 255]$.

(d) The units of the coefficient on nbath is \$1,000 per number of bathrooms.

(e) The units of the intercept are \$1,000. The units of the standard deviation are \$1,000.

13 Profit, Risk, and RD (MLR)

(a) It seems like there is some relationship, especially between RD and profit.



- (b) The plug-in predictive interval, when $RD = 76$ and $RISK = 7$ is $94.75 \pm 2 * 14.34 = [66.1, 123.4]$.

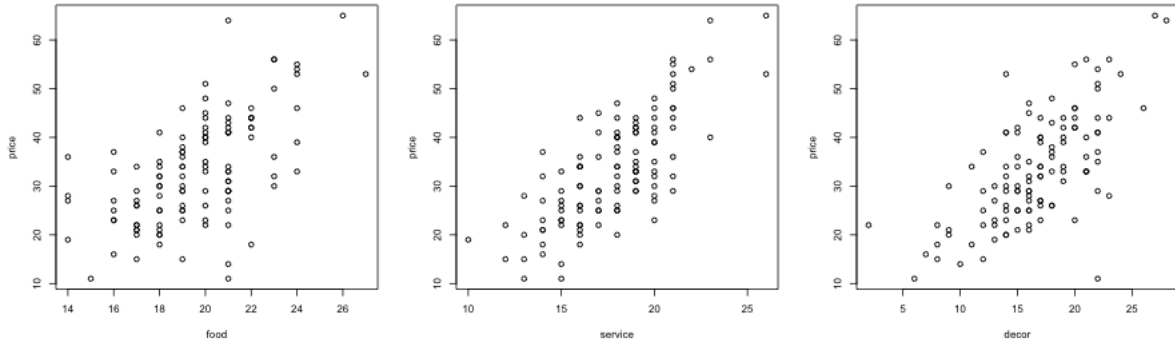
The correct prediction interval, according to R, is $[62.8, 126.7]$.

We know that the plug-in prediction interval is a little too optimistic since it does not take into account any of the uncertainty in the estimates of the parameters.

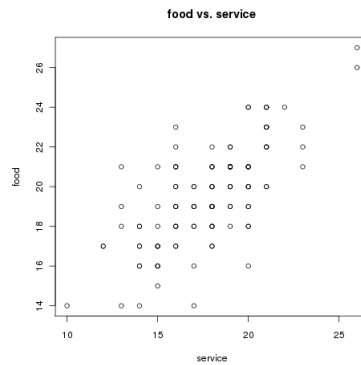
- (c) Using the model $PROFIT = \beta_0 + \beta_1 RISK + \epsilon$, the 68% plug-in prediction interval for when $RISK = 7$ is $143 \pm 106.1 = [37.5, 249.7]$.
- (d) Our interval in (c) is bigger than the interval in (b) despite the fact that it is a “weaker” confidence interval. In essence (c) says that we predict Y will be in $[38, 250]$ 68% of the time when $RISK = 7$. In contrast, (b) says that Y will be in $[63, 127]$ 95% of the time when $RISK = 7$ and $RD = 76$. Using RD in our regression narrows our prediction interval by quite a bit.

14 Zagat (MLR)

- (a) There appears to be some linear relationship between price and food, service, and decor.



- (b) According to R, the 95% predictive interval when food=18, service=14, and decor=16 is [13.8, 39.2].
- (c) The interpretation of the coefficient, $b_1 = 1.380$, on food is that, fixing everything else, a 1 unit increase in food results in a \$1.38 increase in the price of a meal. The regression output is below.
- (d) There is a relationship between food and service. It appears that food and service are positively correlated.



Let us forget about decor for the moment and just think about the regression of involving food *and* service,

$$price = \beta_0 + \beta_1 food + \beta_2 service + error.$$

and the regression *just* involving food,

$$price = \alpha_0 + \alpha_1 food + error.$$

We think that food and service are positively correlated. For simplicity, let's assume that they are perfectly positively correlated. Then those two variables are linearly related like $service = M * food + I$, where $M > 0$. If we replace $service$ in the first regression using this relationship we get

$$price = \beta_0 + \beta_2 I + (\beta_1 + M\beta_2) food + error.$$

Thus if β_2 is positive then $\alpha_1 = \beta_1 + M\beta_2 > \beta_1$. We assume something similar happens even when food and service are not perfectly correlated and we use the full regression.

Following this line of reasoning, the coefficient on service appears to be positive ($b_2 = 1.048$ from the multiple linear regression). Thus for the simple linear regression, $price = \alpha_0 + \alpha_1 food + \epsilon$, we expect the coefficient on food to be higher than the coefficient on food we found when doing the multiple linear regression.

This is indeed the case as the coefficient on food in the multiple regression is 1.38 whereas the coefficient on food in the simple regression is 2.63.

```
lm(formula = price ~ food + service + decor)

Residuals:
    Min       1Q   Median       3Q      Max
-19.0442 -4.0530  0.2109  4.6547 13.0864

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -30.6640    4.7872  -6.405 3.82e-09 ***
food         1.3795    0.3533   3.904 0.000163 ***
service      1.0480    0.3811   2.750 0.006969 **
decor        1.1043    0.1761   6.272 7.18e-09 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 6.298 on 110 degrees of freedom
Multiple R-squared: 0.6875,    Adjusted R-squared: 0.6789
F-statistic: 80.66 on 3 and 110 DF,  p-value: < 2.2e-16

lm(formula = price ~ food)

Residuals:
    Min       1Q   Median       3Q      Max
-25.9774 -6.1653  0.4617  5.8359 27.0226

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.1538    6.5527  -2.77 0.00656 **
food         2.6253    0.3315   7.92 1.89e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 8.939 on 112 degrees of freedom
Multiple R-squared: 0.359,    Adjusted R-squared: 0.3533
F-statistic: 62.72 on 1 and 112 DF,  p-value: 1.893e-12
```

- (e) I would feel less confident prediction the price of a meal when food=20, service=3, and decor=17 because we do not have any data when service is around 3. Using R, the prediction interval given these values is [1.86, 35.8] which is larger than the prediction interval we found in part (b).

15 Salary (MLR)

- (1) The null hypothesis for the F-test is that $\beta_1 = \dots = \beta_p = 0$. The alternative hypothesis is that $\beta_i \neq 0$ for at least one of the i s in $1, \dots, p$.

In this experiment the value of the F-test is 22.98 which is much greater than 4 so we reject the null hypothesis—at least one of the β_i 's, $i \geq 1$ is not zero.

- (2) There does appear to be a positive linear relationship between Salary and Experience, all else equal, since the estimated coefficient on Experience is 1.269 and the t-test is greater than 2. That is the confidence interval for the coefficient on Experience is positive.
- (3) The forecast is

$$33526.4 + 722.5 + 90.02(12) + 1.269(10) + 23.406(15) = 35692.92.$$

- (4) To see if the larger model incorporates useful explanatory variables we use the partial F-test. The partial F-statistic is

$$f = \frac{(R_{full}^2 - R_{base}^2)/(p_{full} - p_{base})}{(1 - R_{full}^2)/(n - p_{full} - 1)}.$$

We can calculate R^2 by $R^2 = SSR/SST$. Doing these calculations we get $f = 20.88$. Again this is much greater, than 4 so we can use our rule of thumb and reject the null hypothesis, which is $\beta_i = 0$ for all the i 's that do not correspond to Education. Rejecting this hypothesis, we conclude that at least one of the other explanatory variables, besides Education, provides useful information.

16 Baseball (Dummy Variables)

- (1) The expected value of R/G given OBP is

$$E[R/G \mid OBP, League = 0] = \beta_0 + \beta_2 OBP$$

for the NL and

$$E[R/G \mid OBP, League = 1] = (\beta_0 + \beta_1) + \beta_2 OBP$$

for the AL.

- (2) β_0 is the number of runs per game we expect a team from the National League to score if their OBP is zero.

We expect a team in the American League to score β_1 more runs per game on average than a team in the National League with the same OBP .

β_2 tells us how R/G scales with OBP . For every unit increase in OBP there will be a β_2 increase in R/G .

- (3) The best guess of β_1 is $b_1 = 0.01615$ with standard error 0.06560. Thus the 99% confidence interval is $b_1 \pm 3 * s_{b_1} = [-0.18, 0.21]$, which includes zero. Since zero is in our interval of reasonable values we cannot conclude that $\beta_1 \neq 0$.

```
lm(formula = R.G ~ League + OBP, data = BB)

Residuals:
    Min       1Q   Median       3Q      Max
-0.23504 -0.10839 -0.05171  0.11360  0.47677

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -7.72065    0.93031  -8.299 6.59e-09 ***
LeagueAmerican  0.01615    0.06560   0.246  0.807
OBP           37.26060    2.72081  13.695 1.14e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.1712 on 27 degrees of freedom
Multiple R-squared:  0.8851,    Adjusted R-squared:  0.8765
F-statistic: 103.9 on 2 and 27 DF,  p-value: 2.073e-13
```

17 Corn Yields (Dummy Variables)

- (a) Creating three dummy variables for fertilizers 1, 2, and 3, our model for crop yields is

$$Y = \beta_0 + \beta_1 F1 + \beta_2 F2 + \beta_3 F3 + \epsilon.$$

(There are four categories here: the three fertilizers and the control group.)

- (b) We want to test that none of the three fertilizers has an effect on crop yields, which is to say that $\beta_1 = \beta_2 = \beta_3 = 0$. To test that these three quantities are zero we use the F-test, which is on page 65 of the notes. Following that formula,

$$f = \frac{R^2/p}{(1 - R^2)/(n - p - 1)} = 5.144,$$

for which $P(F_{3,36} > f) = 0.004605$. Hence we reject the null at 5% and say that at least one $\beta_i, i \geq 1$ is not zero.

```
lm(formula = Y ~ F1 + F2 + F3, data = corn.df)

Residuals:
    Min       1Q   Median       3Q      Max
-9.800 -2.825 -0.600  3.125 10.200

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  29.800    1.533  19.442 < 2e-16 ***
F1           6.800    2.168   3.137 0.00340 **
F2           0.100    2.168   0.046 0.96346
F3           5.100    2.168   2.353 0.02422 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 4.847 on 36 degrees of freedom
Multiple R-squared:  0.3001,    Adjusted R-squared:  0.2417
F-statistic: 5.144 on 3 and 36 DF,  p-value: 0.004605
```

- (c) To test that using fertilizer, any of the three, has a positive effect we create a dummy variable *ANY* which is 1 if fertilizer is used and then run the regression

$$Y = \beta_0 + \beta_1 ANY + \epsilon.$$

At the 5% significance level fertilizer we can (barely) conclude that there is a significant effect. In particular, the 95% confidence interval for β_1 is $[0.092, 7.91]$, which doesn't include zero.

```
Call:
lm(formula = Y ~ ANY, data = corn.df)

Residuals:
    Min     1Q  Median     3Q     Max
-12.80  -3.05   0.20   3.20  11.20

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  29.800      1.692  17.608 <2e-16 ***
ANY           4.000      1.954   2.047  0.0476 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.352 on 38 degrees of freedom
Multiple R-squared:  0.0993,    Adjusted R-squared:  0.0756
F-statistic:  4.19 on 1 and 38 DF,  p-value: 0.04763
```

- (d) From the regression in (a), the largest value of β_i is $\beta_1 = 6.8$. Since all of the β_i 's have the same standard error it seems reasonable to say that fertilizer one is the best.

To make sure though we should run a different regression, this time with dummy variables $F2$, $F3$, and CG , where CG is 1 if Y_i is in the control group and zero otherwise.

```
lm(formula = Y ~ F2 + F3 + CG, data = blah)

Residuals:
    Min     1Q  Median     3Q     Max
-9.800 -2.825 -0.600  3.125 10.200

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  36.600      1.533  23.878 < 2e-16 ***
F2           -6.700      2.168  -3.091  0.00384 **
F3           -1.700      2.168  -0.784  0.43803
CG           -6.800      2.168  -3.137  0.00340 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.847 on 36 degrees of freedom
Multiple R-squared:  0.3001,    Adjusted R-squared:  0.2417
F-statistic:  5.144 on 3 and 36 DF,  p-value: 0.004605
```

We can see from the t -value for $F3$ that we cannot say with 95% probability that fertilizer 1 is better than fertilizer 3. Thus while we are somewhat confident that 1 is better than 3, we are not 95% sure.

18 Mid-City Housing Data (MLR, Dummy Variables)

There may be more than one way to answer these questions.

- (1) To begin we create dummy variable *Brick* to indicate if a house is made of brick and N_2 and N_3 to indicate if a house came from neighborhood two and neighborhood three

respectively. Using these dummy variables and the other covariates, we ran a regression for the model

$$Y = \beta_0 + \beta_1 \text{Brick} + \beta_2 N_2 + \beta_3 N_3 + \beta_4 \text{Bids} + \beta_5 \text{SqFt} + \beta_6 \text{Bed} + \beta_7 \text{Bath} + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2).$$

and got the following regression output.

```
lm(formula = Price ~ Brick + N2 + N3 + Offers + SqFt + Bedrooms +
    Bathrooms, data = house)

Residuals:
    Min       1Q   Median       3Q      Max
-27337.3  -6549.5   -41.7   5803.4  27359.3

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2159.498    8877.810     0.243  0.808230
BrickYes     17297.350    1981.616     8.729  1.78e-14 ***
N2          -1560.579    2396.765    -0.651  0.516215
N3          20681.037    3148.954     6.568  1.38e-09 ***
Offers      -8267.488    1084.777    -7.621  6.47e-12 ***
SqFt         52.994         5.734     9.242  1.10e-15 ***
Bedrooms    4246.794    1597.911     2.658  0.008939 **
Bathrooms   7883.278    2117.035     3.724  0.000300 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 10020 on 120 degrees of freedom
Multiple R-squared:  0.8686,    Adjusted R-squared:  0.861
F-statistic: 113.3 on 7 and 120 DF,  p-value: < 2.2e-16
```

To check if there is a premium for brick houses given everything else being equal we test the hypothesis that $\beta_1 = 0$ at the 95% confidence level. Using the regression output we see that the 95% confidence interval for β_1 is [13373.89, 21220.91]. Since this does not include zero we conclude that brick is a significant factor when pricing a house. Further, since the entire confidence interval is greater than zero we conclude that people pay a premium for a brick house.

- (2) To check that there is a premium for houses in Neighborhood three, given everything else we repeat the procedure from part (1), this time looking at β_3 . The regression output tells us that the confidence interval for β_3 is [14446.33, 26915.75]. Since the entire confidence interval is greater than zero we conclude that people pay a premium to live in neighborhood three.
- (4) We want to determine if Neighborhood 2 plays a significant role in the pricing of a house. If it does not, then it will be reasonable to combine neighborhoods one and two into one “old” neighborhood. To check if Neighborhood 2 is important, we perform a hypothesis test on $\beta_2 = 0$. The null hypothesis $\beta_2 = 0$ corresponds to the dummy variable N_2 being unimportant. Looking at the confidence interval from the regression output we see that the 95% confidence interval for β_2 is [-6306, 3184], which includes zero. Thus we can conclude that it is reasonable to let β_2 be zero and that neighborhood 2 may be combined with neighborhood 1.
- (3) To check that there is a premium for brick houses in neighborhood three we need to alter our model slightly. In particular, we need to add an interaction term $\text{Brick} \times N_3$.

This more complicated model is

$$Y = \beta_0 + \beta_1 \text{Brick} + \beta_2 N_2 + \beta_3 N_3 + \beta_4 \text{Bids} + \beta_5 \text{SqFt} + \beta_6 \text{Bed} + \beta_7 \text{Bath} + \beta_8 \text{Brick} \cdot N_3 + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2).$$

To see what this interaction term does, observe that

$$\frac{\partial E[Y|\text{Brick}, N_3]}{\partial N_3} = \beta_3 + \beta_8 \text{Brick}.$$

Thus if β_8 is non-zero we can conclude that consumers pay a premium to buy a brick house when shopping in neighborhood three. The output of the regression which includes the interaction term is below.

```
lm(formula = Price ~ Brick + N2 + N3 + Offers + SqFt + Bedrooms +
    Bathrooms + Brick * N3, data = house)

Residuals:
    Min       1Q   Median       3Q      Max
-26939.1 -5428.7  -213.9   4519.3 26211.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3009.993    8706.264     0.346  0.73016
BrickYes     13826.465    2405.556     5.748 7.11e-08 ***
N2           -673.028    2376.477    -0.283  0.77751
N3           17241.413    3391.347     5.084 1.39e-06 ***
Offers       -8401.088    1064.370    -7.893 1.62e-12 ***
SqFt         54.065       5.636      9.593 < 2e-16 ***
Bedrooms     4718.163    1577.613     2.991 0.00338 **
Bathrooms    6463.365    2154.264     3.000 0.00329 **
BrickYes:N3  10181.577    4165.274     2.444 0.01598 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 9817 on 119 degrees of freedom
Multiple R-squared:  0.8749,    Adjusted R-squared:  0.8665
F-statistic:  104 on 8 and 119 DF,  p-value: < 2.2e-16
```

		0.5 %	99.5 %
(Intercept)	-19781.05615	25801.04303	
BrickYes	7529.25747	20123.67244	
N2	-6894.11333	5548.05681	
N3	8363.62557	26119.20030	
Offers	-11187.37034	-5614.80551	
SqFt	39.31099	68.81858	
Bedrooms	588.32720	8847.99967	
Bathrooms	823.98555	12102.74436	
BrickYes:N3	-722.17781	21085.33248	

To see if there is a premium for brick houses in neighborhood three we check that the 95% confidence interval is greater than zero. Indeed, we calculate that the 95% confidence interval is [1933, 18429]. Hence we conclude that there is a premium at the 95% confidence level. Notice however, that the confidence interval at the 99% includes zero. Thus if one was very stringent about drawing conclusions from statistical data, they may accept the claim that there is no premium for brick houses in neighborhood three.